BAYESIAN ESTIMATION OF CORRELATION MATRICES OF LONGITUDINAL DATA AND VARIABLE CLUSTERING

A Dissertation

by

RIDDHI PRATIM GHOSH

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bani Mallick
Co-Chair of Committee,	Mohsen Pourahmadi
Committee Members,	Anirban Bhattachrya
	Raktim Bhattacharya
Head of Department,	Jianhua Huang

August 2019

Major Subject: Statistics

ABSTRACT

Estimation of correlation matrices is a challenging problem due to the notorious positive-definiteness constraint and high-dimensionality. Reparameterising Cholesky factors of correlation matrices in terms of angles or hyperspherical coordinates where the angles vary freely in the range $[0, \pi)$ has become popular in the last two decades. However, it has not been used in Bayesian estimation of correlation matrices perhaps due to lack of clear statistical relevance and suitable priors for the angles. In this dissertation, we show for the first time that for longitudinal data these angles are the inverse cosine of the semi-partial correlations (SPCs). This simple connection makes it possible to introduce physically meaningful selection and shrinkage priors on the angles or correlation matrices with emphasis on selection (sparsity) and shrinking towards special structures. Our method deals effectively with the positive-definiteness constraint in posterior computation. We compare the performance of our Bayesian estimation based on angles with some recent methods based on partial autocorrelations through simulation and apply the method to data related to clinical trial on smoking. Subsequently this reparametrization has been exploited in a variable clustering problem which focuses on model-based clustering of components of a k-dimensional random vector hinging on a block diagonal correlation structure with equicorrelated blocks. There are plenty of data-driven and model based clustering algorithms available in the literature for data clustering. However, literature on variable clustering is limited. We adopt a model-based approach for variable clustering which assumes an inherent probabilistic model determining the clusters. Starting from a multivariate normal likelihood, we enforce the clustering through prior modeling. With unknown number of clusters, we assume a truncated Poisson distribution (by penalizing large number of clusters) as prior for number of clusters and perform a reversible jump Markov Chain Monte Carlo to correctly estimate the number of clusters in the posterior computation. The end product of our algorithm is cluster recovery of the variables along with the estimation of number of clusters. The performance of the algorithm has been substantiated with extensive simulation studies and a real data example from genetics.

DEDICATION

To my mother, my father, grandparents and relatives.

ACKNOWLEDGMENTS

First and foremost, I acknowledge my parents for being extremely supportive of my education throughout and lots of sacrifices they made to make my life smooth. Their liberal outlook and interest in academia have always been an inspiration, and they continue to support and motivate me in all my endeavors. My heartfelt gratitude goes out to my doctoral advisors, Prof. Mohsen Pourahmadi and Professor Bani Mallick for being supportive in the journey of last five years, and Prof. Anirban Bhattacharya and Prof. Debdeep Pati for their valuable guidance and cooperation in various aspects. An especially profound thanks to Prof. Pourahmadi for being a great mentor. I would like to take this opportunity to thank all my teachers from primary and high school, my professors from bachelors and master's programmes at the Indian Statistical Institute, Kolkata and my professors at the Texas A&M University (notably Prof. Valen Johnson). I would like to express my thanks to my close friends (notably Bikram, Pritam) whose constant supply of courage and persistence was one of the key ingredients in this journey.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Mohsen Pourahmadi, Bani Mallick and Anirban Bhattacharya of the Department of Statistics and Professor Raktim Bhattacharya of the Department of Aerospace Engineering.

The data analyzed for Chapter 3 was provided by Professor Bani Mallick. All work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and partially by Cloud Computing Based Robust Space Situational Awareness grant from AFOSR.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

In this chapter, we present a thorough literature review of Bayesian analysis of covariance and correlation matrices. The subject is vast and growing vary rapidly. In the regression based approach to covariance estimation, we consider priors for a covariance matrix introduced through regression parameters M. J. Daniels & Pourahmadi (2002), Smith & Kohn (2002) and Fox & Dunson (2011). Furthermore, it is instructive to note that in a regularized regression set-up, the penalty term $p(\beta)$ when exponentiating to exp $(-p(\beta))$ leads to a prior for regression parameters β Tibshirani (1996).

It is interesting to note that the starting point of modern trend of prior elicitation for covariance matrices is various matrix decomposition which we discuss more in the subsequent subsections in this chapter. In the early development of Bayesian covariance estimation, traditional Jeffreys' improper prior and the conjugate inverse Wishart (IW) priors were in practice in the works of Lin (1985), P. J. Brown et al. (1994), due to their conjugacy.

Later in 1980s, the success of Markov chain Monte Carlo (MCMC) based computation facilitated the possibility of flexible and novel priors which went beyond the traditional Jeffyes' or IW priors. For more details, we refer the readers to Yang & Berger (1994), M. J. Daniels & Kass (2001), Wong et al. (2003), Hoff (2009). Some of these priors were inspired by certain features of IW prior and gave rise to the generalized inverse Wishart (GIW) prior introduced by P. J. Brown et al. (1994), M. Daniels & Pourahmadi (2002), Smith & Kohn (2002), Barnard et al. (2000), which rely either on the Cholesky decomposition or variance-correlation decomposition (*separation strategy*). In the next few subsections, we present the related work in Bayesian covariance estimation in chronological order.

1.1 Priors via Spectral Decomposition

Since the seminal work of Stein (1956), estimation of covariance matrix has led shrinking the eigenvalues of the sample covariance matrix to a common value, see Dey et al. (1985), Lin (1985), Yang & Berger (1994), M. J. Daniels & Kass (1999), Hoff (2009). Such estimators have lower risk

than sample covariance matrix. Shrinking eigenvectors have been shown to have lower estimated risk (M. J. Daniels & Kass (1999), Johnstone & Lu (2009)).

There are broadly three classes of priors which are based on unconstrained parameterization of a covariance matrix using its spectral decomposition, with the objective of shrinking some functions of the off-diagonal elements of covariance or correlation matrix to a common value. This results in estimating smaller number of parameters as compared to $k(k-1)/2$ dependent parameters for a k-dimensional covariance matrix.

The log matrix prior of Leonard et al. (1992) uses the matrix logarithm of the covariance matrix Σ, defined as,

$$\log \Sigma = P(\log \Lambda)P^\top,$$

where $\Sigma = P\Lambda P^\top$ is the spectral decomposition of Σ and $\log \Lambda = \text{ diag}(\log \lambda_1, \log \lambda_1, \cdots, \log \lambda_k)^\top$.

Multivariate normal prior on the entries of $\log \Sigma$ has been used by Leonard et al. (1992). The advantages of this prior are easily understood for the covariance matrix of a multivariate normal distribution providing a hierarchical and empirical Bayes inference compared to the conjugate IW prior, which lacks such flexibility. P. J. Brown et al. (1994) aptly pointed out that this prior lacks statistical interpretability of the elements of $\log \Sigma$. Also the relationship between entries of $\log \Sigma$ and Σ are highly complicated. Due to the lack of interpretability, choice of hyperparameters leads to difficulty.

The reference prior of Yang & Berger (1994) is of the form,

$$p(\Sigma) = c[|\Sigma| \prod_{i<j}(\lambda_i - \lambda_j)]^{-1},$$

where $\lambda_1 > \lambda_2 > ... > \lambda_k$ are the ordered eigenvalues of Σ and c is a constant. It is known (Yang & Berger (1994), p. 1194) that compared to Jeffreys prior, the reference prior puts more mass near the region of equality of eigenvalues. This makes the reference prior effective in producing estimators with better eigenstructure shrinkage. It is interesting to note that the reference prior

for Σ^{-1} and the eigenvalues of the covariance matrix are the same as $p(\Sigma)$. Expression for Bayes estimate of covariance matrix involves computation of high-dimensional posterior expectations, where the computation is carried out through the hit-and-run sampler in MCMC setup.

Interestingly, M. J. Daniels (2005) has shown that the reference prior of Yang & Berger (1994) corresponds to a uniform prior on the matrix P and flat improper priors on the logarithms of the eigenvalues of Σ. The shrinkage prior of M. J. Daniels & Kass (1999) also rely on the spectral decomposition of the covariance matrix and shrinks the eigenvectors by reparameterizing the orthogonal matrix in terms of $k(k-1)/2$ Givens angles Computation et al. (1996) θ between pairs of columns of orthogonal matrix P. Since θ lies in $(-\pi/2, \pi/2)$, with a logit transform one may use a mean zero normal prior on them. However, the statistical interpretation of Givens angles as parameters is not well understood. The idea of using matrix Bingham distributions as priors on the group of orthogonal matrices Hoff (2009) is a major recent contribution to eigenvectors shrinkage of the sample covariance matrix.

Using simulations, Yang & Berger (1994) compared the performance of reference prior to the covariance estimator of Haff et al. (1991) and found it competitive corresponding to certain loss functions. M. J. Daniels & Kass (2001) provided simulation based performance of their shrinkage estimator compared to other Bayes estimator in terms of Stein's loss function. They found that the estimators of reference prior of Yang & Berger (1994) performs as good as those based on Givens angles for some non-diagonal and ill-conditioned matrices, whereas underperforms when the true covariance matrix is diagonal and poorly conditioned.

1.2 Inverse-Wishart (IW) and Generalized Inverse-Wishart (GIW) priors

In this subsection, we review inverse Wishart (IW) prior and some of its extension.

1.2.1 Inverse-Wishart (IW) prior

In the Bayesian estimation of a covariance, inverse-Wishart prior has been used extensively West & Escobar (1993),Barnard et al. (2000), Bernardo & Smith (2001) as it is the natural conjugate prior for normal model and thus, serves as a prior on the residual covariance matrix in

multivariate regression model Box & Tiao (2011). It is of the form

$$p(\Sigma) \propto |\Sigma|^{-(\nu+k+1)/2} \exp\{-\text{tr}(\Psi\Sigma^{-1})/2\}$$

with scale matrix Ψ and degrees of freedom parameter ν Wishart (1928); Press (1982). However, single degrees of freedom parameter restricts this prior from eliciting substantive prior information about degrees of correlation among variables Gelman et al. (2014). The other priors involve Jeffreys prior, $p(\Sigma) \propto |\Sigma|^{-(k+1)/2}$; the log matrix prior Leonard et al. (1992) using logarithmic transformation of eigen decomposition of Σ ; reference prior Yang & Berger (1994) of the form $p(\Sigma) \propto 1/\{|\Sigma| \prod_{i<j}(d_i - d_j)\}$, where d_i are the eigenvalues of Σ. However, for the priors parameterized in the similar fashion as in later two models, non-linearity of the relationship between (log) eigenvalues and correlations makes the interpretation of the new parameters arduous.

1.2.2 Generalized Inverse-Wishart (GIW) prior

Hierarchical extension of the inverse-Wishart prior in M. J. Daniels & Kass (1999) is bit more flexible than traditional use of inverse-Wishart prior by introducing priors on the degrees of freedom parameter and the diagonal elements of the scale matrix Ψ. They proposed flat prior on the logarithms of the diagonal elements of the Wishart scale matrix and a flat prior on the logarithm of the degrees of freedom parameter trunacted at a large value. We note that if the scale matrix Ψ has elements $(1/a_1, 1/a_2, \cdots, 1/a_k)$ on its diagonal, then $p(\log(1/a_j)) \propto c$, a flat prior. The distribution of individual a_j is $p(a_j) \propto c/a_j$ which is an improper prior for the individual a_js for $j = 1, 2, \cdots, k$. They have proposed another prior based on Givens angles. Writing the spectral decomposition of $\Sigma = P\Lambda P^\top$, where Λ is a diagonal matrix having ordered eigenvalues, one can express $P = G_{12}G_{13}\cdots G_{1,k-1}G_{k-1,k}$, where G_{ij} is the $k \times k$ identity matrix with the i^{th} and j^{th} diagonal elements replaced by $\cos\theta_{ij}$ and the (i, j) and (j, i) elements replaced by $\pm \sin\theta_{ij}$. They have put a normal distribution on a logit transformation of the angles; $\log([\pi/(2+\theta)]/[\pi/(2-\theta)]) \sim N(0, \tau^2)$ and use $p(\tau^2) \sim (c + \tau^2)^{-1}$ and flat priors on eigen-

values. In comparison to to the extension of hierarchical inverse-Wishart prior, Jeffreys prior and Berger's reference prior Yang & Berger (1994), the later prior performed the best with respect to the Kullback-Liebler loss function.

Following Gelman et al. (2006), Huang et al. (2013) proposes inverse-Wishart distribution for the covariance matrix assuming $\Psi = \text{diag}(1/a_1, 1/a_2, \cdots, 1/a_k)$ to be a diagonal matrix and fixing degrees of freedom parameter ν as,

$$\begin{aligned}
&\Sigma | a_1, a_2, \cdots, a_k \sim \text{Inverse-Wishart}(\nu + k - 1, 2\nu \ \text{diag}(1/a_1, 1/a_2, \cdots, 1/a_k)), \\
&a_j \sim \text{Inverse-Gamma}(1/2, 1/A_j^2) \text{ for } j = 1, 2, \cdots, k,
\end{aligned} \tag{1.1}$$

where ν, A_js are positive scalars. The upshots of this prior distribution are:
(i) The marginal distribution of any sub-covariance matrix has Inverse-Wishart distribution.
(ii) The marginal distribution of any standard deviation σ_j in Σ is Half-$t(\nu, A_j)$ distribution which is a non-informative prior.
(iii) The marginal distribution of correlations r_{ij} in Σ is of the form

$$p(r_{ij}) \propto (1 - r_{ij}^2)^{\nu/2-1}, -1 \leq r_{ij} \leq 1. \tag{1.2}$$

which is an extended Beta distribution on $(-1, 1)$.
Another hierarchical extension of the inverse-Wishart prior is due to P. Brown (2002) to circumvent the issue pointed out by Gelman et al. (2014), i.e. to control the uncertainly of $k(k-1)/2$ parameters by a single degrees of freedom parameter ν. In this context, P. Brown (2002) defined generalized inverted Wishart distribution(GIW) by partitioning k variables into b blocks.
Let $\{q_i\}_{i=1}^b$ be the partition with $q = \sum_{i=1}^b q_i$. Define

$$q_{\{i\}} = \sum_{j=1}^{i} q_j.$$

The covariance matrix Σ is partitioned into $\Sigma = (\Sigma_{ij})$ where Σ_{ij} is a $q_i \times q_j$ submatrix. Define the sequence for $i = 1, 2, \cdots, (b-1)$

$$B_i = \Sigma_{\{i\}\{i\}}^{-1} \Sigma_{\{i\}(i+1)} \tag{1.3}$$

$$\Gamma_i = \Sigma_{(i+1)(i+1)} - \Sigma_{(i+1)\{i\}} \Sigma_{\{i\}\{i\}}^{-1} \Sigma_{\{i\}(i+1)} \tag{1.4}$$

With the above set-up, Generalized Inverted Wishart(GIW) distribution is defined as

Definition 1.2.1. *Let the q variables be partitioned into b sets. Let Σ_{11} and (B_i, Γ_i), for $i = 1, 2, \cdots, (b-1)$ be mutually independent. Assume the matrices B_{*i}, Q_i, H_i and the scalars δ_i, $i = 1, 2, \cdots, b-1$ are constants and also δ_0, Q_0 are constants. If*

$$\Sigma_{11} \sim IW(\delta_0, Q_0) \tag{1.5}$$

$$B_i | \Gamma_i \sim B_{*i} + N(H_i, \Gamma_i) \tag{1.6}$$

$$\Gamma_i \sim IW(\delta_i, q_{\{i\}} Q_i) \tag{1.7}$$

Then Σ follows a GIW distribution with paramaters $(\delta_0, \delta_i; B_{*i}; H_i, Q_0, Q_i; i = 1, 2, \cdots, b-1)$.

The following consequences are immediate.

1. When $b = 1$, GIW reduces to IW distribution for Σ.

2. For $b = q = k$ and $q_j = 1$, this leads to the modified Cholesky decomposition of Σ in which case ((1.5)-(1.7)) reduces to the prior in M. J. Daniels & Pourahmadi (2002).

Denote a matrix normal distribution by $N(.,.)$. In a multivariate regression model:

$$Y_{n \times k} - X_{n \times p} B_{p \times k} \sim N(I_n, \Sigma_{k \times k}) \tag{1.8}$$

P. Brown (2002) discussed four possible ways of getting a GIW posterior distribution, see p.4 P. Brown (2002) for more details.

In the context of dynamic modeling of longitudinal data, M. J. Daniels & Pourahmadi (2002) diagonalizes covariance matrix as $B\Sigma B^\top = D$ and uses Gaussian prior on the unconstrained nonredundant entries of lower-triangular matrix B and inverse-Gamma distribution on the elements of the diagonal matrix D. Inverse-Wishart prior on Σ appears as a special case of suitable choices of the hyper-paramaters in this set-up.

1.2.3 Dynamic Inverse-Wishart (DIW) prior

In multivariate time series, Lan et al. (2017) uses the model

$$y_t \sim N(\mu_t, \Sigma_t)$$

for the time indexed by t and expressing $\Sigma_t = S_t B_t B_t^\top S_t$, they assume Gaussian process priors with exponential covariance kernel for μ_t, logarithm of the elements of diagonal S_t and each row for the lower-triangular matrix B_t. Wilson & Ghahramani (2010) discusses simulation scheme of inverse-Wishart process staring from a Gaussian process.

1.3 Priors on Correlation Matrices

1.3.1 Separation Strategy

The first use of variance-correlation factorization in Bayesian covariance estimation is due to Barnard et al. (2000) using the factorization $p(\Sigma) = p(D, R) = p(D)p(D|R)$ and introducing independent priors for the standard deviations in D and the correlations in R. In particular, they used log-normal priors on the variances independent of a prior on the whole matrix R capable of inducing uniform $(-1, 1)$ priors on its entries r_{ij}. This is done by first deriving the marginal

distribution of R when Σ has a standard IW distribution $W_p^{-1}(I, \nu)$, $\nu \geq k$ with the density

$$f_k(\Sigma|\nu) = c|\Sigma|^{-\frac{1}{2}(\nu+k+1)} \exp\big(-\frac{1}{2}tr\Sigma^{-1}\big).$$

It turns out that

$$f_k(R|\nu) = c|R|^{\frac{1}{2}(\nu-1)(k-1)-1} \prod_{i=1}^{k} |R_{ii}|^{-\nu/2},$$

where R_{ii} is the principle submatrix of R. Then, using the marginalization property of the IW (principle submatrix of an IW is again an IW), the marginal distribution of each r_{ij} is obtained as

$$f(r_{ij}|\nu) = c(1 - r_{ij}^2)^{\frac{\nu-k-1}{2}}, \quad |r_{ij}| \leq 1,$$

which is a Beta($\frac{\nu-k+1}{2}, \frac{\nu-k+1}{2}$) on $(-1, 1)$ and reduces to the uniform distribution when $\nu = k + 1$. Moreover, choosing either $k \leq \nu < k + 1$ or $\nu > k + 1$, one can control the tail of $f(r_{ij}|\nu)$. The above family of priors for R is indexed by a single "tuning" parameter ν.

In the context of Bayesian dynamic modeling of covariance and correlation matrix, Lan et al. (2017) used the fact that correlations can be represented as vectors on unit sphere and proposed the following distribution.

1.3.2 Squared-Dirichlet distribution prior

A random vector $b_l \in S^{l-1}$(l-dimesnional unit sphere) follows a squared Dirichlet distribution if $b_l^2 = (b_{l1}^2, b_{l2}^2, \cdots, b_{ll}^2)^\top \sim$ Dir $(\alpha_{l1}, \alpha_{l2}, \cdots, \alpha_{ll})$. Denoting $b_l \sim$ Dir$^2(\alpha_l)$, where $\alpha_l = (\alpha_{l1}, \alpha_{l2}, \cdots, \alpha_{ll})^\top$, it is given in Lan et al (2017) that the corresponding distribution of b_l is

$$p(b_l) = p(b_{l1}, b_{l2}, \cdots, b_{ll}) \propto \prod_{u=1}^{l} |b_{lu}|^{2\alpha_{lu}-1} \tag{1.9}$$

In a Cholesky based approach by writing $R = BB^\top$, where b_ls are the rows of the lower triangular matrix B , they used the aforementioned prior for Bayesian dynamic modeling of correlation

matrices.

2.1 Reparameterizations of Correlation Matrix

Unconstrained parameterization of correlation matrices using angles and partial autocorrelations has been around for a while (Pinheiro & Bates (1996), Rapisarda et al. (2007), Joe (2006), Madar (2015)) where the partial autocorrelations and angles as new parameters vary freely in the ranges$[-1, 1]$ (Joe (2006)) and $[0, \pi)$ (Pinheiro & Bates (1996), Pourahmadi & Wang (2015), Tsay & Pourahmadi (2017)), respectively. These are described in the following few sections.

2.1.1 PACF based reparameterization

A k-dimensional correlation matrix R with 1s on its diagonal can be reparameterized in terms of the correlations $\rho_{i,i+1}$ for $i = 1, 2, \cdots, k-1$ and partial correlations $\rho_{j,j+l|j+1,...,j+l-1}$ for $j = 1, 2, ..., k-l$ and $l = 2, ..., k-1$, where the formula for computing $\rho_{j,j+l|j+1,...,j+l-1}(2 \leq l \leq k-1)$ is given in Anderson (2003),

$$\frac{\rho_{j,j+l} - r_1^\top(j,l)(R_2(j,l))^{-1}r_3(j,l)}{\big(1 - r_1^\top(j,l)(R_2(j,l))^{-1}r_1(j,l)\big)\big(1 - r_3^\top(j,l)(R_2(j,l))^{-1}r_3(j,l)\big)},$$

where $r_1(j,l) = (\rho_{j,j+1}, ..., \rho_{j,j+l-1})^\top$, $r_3 = (\rho_{j+l,j+1}, ..., \rho_{j+l,j+l-1})^\top$ and $R_2(j,l)$ is the correlation matrix corresponding to the components $(j+1, ..., j+l-1)$. Note that relation between correlations and the partial correlations are indeed invertible.

2.1.2 Semi-partial correlation based reparameterization

There are two alternative parameterizations of a correlation matrix (Madar (2015)) where the entries of the Cholesky factor are expressed in terms of semi-partial correlations and the successive Schur-complements of R. These two are summarized in the following two lemmas. Let the *semi-partial correlation coefficients* $\rho_{ij(1,2,...,i-1)}$ be defined as

$$\rho_{ij(1,2,...,i-1)} = \frac{\rho_{ij} - \rho_i^j R_{i-1}^{-1}\rho_i}{\sqrt{1 - \rho_i R_{i-1}^{-1}\rho_i^\top}} \tag{2.1}$$

where $\rho_i = \rho_i^i = (r_{1i}, r_{2i}, ..., r_{i-1,i})$.

Lemma 2.1.1. *For a correlation matrix $R = (r_{ij})$, let $\rho_i^j = (r_{1j}, r_{2j}, ..., r_{i-1,j})$ for $j \geq i$. Then the lower Cholesky factor $B = (b_{ij})$ of R is given by*

$$b_{ji} = \begin{cases} \rho_{ij(1,2,...,i-1)} & \text{if } i < j, \\ \sqrt{1 - \rho_i R_{i-1}^{-1} \rho_i^\top} & \text{if } i = j, \end{cases} \tag{2.2}$$

where R_{i-1}^{-1} is the inverse of the matrix $R_{i-1} = (r_{kj})_{k,j=1}^{i-1}$.

Lemma 2.1.2. *Let $s_{ij} = sign(\rho_{ij(1,2,...,i-1)})$ and define R_i^j as*

$$R_i^j = \begin{bmatrix} R_{i-1} & \rho_i^{j\top} \\ \rho_i^j & 1 \end{bmatrix} \tag{2.3}$$

Then the lower triangular Cholesky factor B of R can be written as

$$B = \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ s_{12}\sqrt{1 - |R_2|} & \sqrt{|R_2|} & 0 & ... & 0 \\ s_{13}\sqrt{1 - |R_2^3|} & s_{23}\sqrt{|R_2^3| - \frac{|R_3|}{|R_2|}} & \sqrt{\frac{|R_3|}{|R_2|}} & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{1p}\sqrt{1 - |R_2^p|} & s_{2p}\sqrt{|R_2^p| - \frac{|R_3^p|}{|R_2|}} & s_{3p}\sqrt{\frac{|R_3^p|}{|R_2|} - \frac{|R_4^p|}{|R_3|}} & ... & \frac{|R_p|}{|R_{p-1}|} \end{bmatrix} \tag{2.4}$$

2.2 Angle based reparameterization

This section describes a connection between the well-known hyperspherical coordinates (angles) of the Cholesky factor of a correlation matrix $R = (r_{ij})$ and the less familiar semi-partial correlation coefficients $\rho_{ji:1,2,...,j-1}$ between the variables y_i and y_j $(i > j)$ conditioned on the previous variables $y_1, y_2, ..., y_{j-1}$, see Huber (1981), Eaves & Chang (1992), and Madar (2015).

For a general $k \times k$ correlation matrix R with 1's on the diagonal, its Cholesky decomposition is given by $R = BB^\top$ where the Cholesky factor B is a lower triangular matrix. Since the rows

of B are vectors of unit-length, it turns out that they admit the following representation involving trigonometric functions of some angles (Pinheiro & Bates, 1996; Rapisarda et al., 2007):

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ c_{21} & s_{21} & 0 & 0 & \dots & 0 \\ c_{31} & c_{32}s_{31} & s_{32}s_{31} & 0 & \dots & 0 \\ c_{41} & c_{42}s_{41} & c_{43}s_{42}s_{41} & \prod_{j=1}^{3} s_{4j} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ c_{k1} & c_{k2}s_{k1} & c_{k3}s_{k2}s_{k1} & c_{k4}\prod_{j=1}^{3} s_{kj} & \dots & \prod_{j=1}^{k-1} s_{kj} \end{bmatrix} \tag{2.5}$$

with $c_{ij} = \cos(\theta_{ij})$ and $s_{ij} = \sin(\theta_{ij})$, where the angles θ_{ij}'s are measured in radians, $1 \leq j < i \leq k$. Restricting $\theta_{ij} \in [0, \pi)$ makes the diagonal entries of B non-negative, and hence B is unique to which we can associate a $(k-1) \times (k-1)$ lower triangular matrix Θ with $k(k-1)/2$ angles:

$$\Theta = \begin{bmatrix} \theta_{21} & 0 & 0 & \dots & 0 \\ \theta_{31} & \theta_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \theta_{k2} & \theta_{k3} & \dots & \theta_{k,k-1} \end{bmatrix}$$

Note that the (i, j)-th element of Θ is denoted by $\theta_{i+1,j}$ so that θ_{ij} corresponds to the (i, j)-th element of R, we refer to Θ as the **angular matrix** associated to R. For further details, properties and applications of these angles, see Creal et al. (2011), Zhang et al. (2015) and Tsay & Pourahmadi (2017).

In fact, one may go back and forth between R and $\Theta(B)$ using the following forward (T_f) and backward (T_b) transformations $T_f : R \to \Theta$ and $T_b : \Theta \to R$, respectively, as described next.

The transformation T_f: Given a correlation matrix R (symmetric and positive definite) and its Cholesky decomposition $R = BB^\top$ with entries b_{ij}, matching entries of both sides it follows

that

$$b_{11} = 1, \;\; b_{i1} = r_{i1}, \;\; i = 2, \ldots, k. \tag{2.6}$$

Thus, the entries in the first columns of B and R are the same. The θ_{ij}s, the entries of Θ are computed recursively via

The transformation T_b: Given the matrix Θ with entries $\theta_{ij} \in [0, \pi)$, construct the lower triangular matrix $B = (b_{ij})_{i,j \in \{1,2,\ldots,k\}}$ via

$$b_{ij} = \begin{cases} \prod_{l=1}^{i-1} \sin(\theta_{il}), \text{ for } i = j \\ \cos(\theta_{ij}) \prod_{l=1}^{j-1} \sin(\theta_{il}), \text{ for } 1 \leq j < i \leq k, \end{cases} \tag{2.7}$$

and then the correlation matrix $R = BB^\top$.

2.2.1 Examples

We illustrate the nature of nonlinear relation between the angles and correlations using the following three simple examples of increasing dimensions:

(a) For $k = 2$, there is only one angle θ_{21} and

$$B = \begin{bmatrix} 1 & 0 \\ \cos\theta_{21} & \sin\theta_{21} \end{bmatrix},$$

and from (2.6) we obtain the simple relation $r_{21} = \cos\theta_{21}$. Then, the statistical meaning of θ_{21} as the inverse cosine of r_{21} is fairly clear.

(b) For $k = 3$, we have

$$\Theta = \begin{bmatrix} \theta_{21} & 0 \\ \theta_{31} & \theta_{32} \end{bmatrix}.$$

The relationship between r_{ij}'s and θ_{ij}'s are

$$\theta_{21} = \arccos(r_{21}), \theta_{31} = \arccos(r_{31}), \theta_{32} = \arccos\Big(\frac{r_{32} - \cos(\theta_{21})\cos(\theta_{31})}{\sin(\theta_{21})\sin(\theta_{31})}\Big), \tag{2.8}$$

$$\tag{2.9}$$

so that while the two angles in the first column are tied to the individual marginal correlations as in (2.6), this is not the case for θ_{32} in the second column. In fact, the situation gets more complicated for larger k's. In general, the entries of the first column of Θ are just the inverse cosine of the respective entries of the first column of R, but as one moves towards its last column the expression for θ_{ij} becomes more complicated and hence less interpretable as a function of the entries of R.

(c) Block common correlation matrices: As a generalization of a compound symmetric (exchangeable) correlation matrix, consider a block common correlation matrix which is a blocked-matrix where the correlations within each block are equal and different across blocks. Such matrices arise in many applications due presence of common (latent) factors in different regions (aggregation of carbon dioxide sequestration storage assessment units Blondes et al. (2013)) and stock returns of different companies within the same industry (Liechty et al., 2004; Tsay & Pourahmadi, 2017). As an illustration, we consider the following 6×6 correlation matrix,

$$R = \begin{pmatrix} 1 & r_1 & r_2 & r_2 & r_3 & r_3 \\ & & 1 & r_4 & r_5 & r_5 \\ & & & 1 & r_5 & r_5 \\ & & & & 1 & r_6 \\ & & & & & 1 \end{pmatrix},$$

with 6 distinct correlations r_i, $i = 1, 2, ..., 6$, which is much smaller than 15, the number of distinct entries of a generic correlation matrix of this size. The corresponding matrix of angles Θ, is completely determined by six *pivotal angles* denoted by $(\theta_{21}, \theta_{31}, \theta_{51}, \theta_{43}, \theta_{53}, \theta_{65})$ (Tsay & Pourah-

madi, 2017) where their subscripts indicate their locations in the partitioned matrix Θ.

In general, for a $k \times k$ correlation matrix R, if there is d common correlation blocks, then the pivotal angles consist of d angles in the range $[0, \pi)$, say $\theta_{pivotal} = (\theta_1, \theta_2, ..., \theta_d)^\top$. The other angles in Θ matrix, called *implied angles*, are functions of pivotal angles and can be obtained using an algorithm in Tsay & Pourahmadi (2017). When the blocks are known as in the above example, it is simple to determine the positions of the pivotal angles, and this will reduce the dimension of the parameter space to d and hence the computational cost.

2.2.2 The angles and semi-partial correlations

Statistical interpretation and plausible meaning of the angles as the new parameters of a correlation matrix are of interest when eliciting priors. This task is complicated by the nonlinearity of the relationships between the correlations and angles as seen in (2.8).

Here, we use a relatively dormant formula for b_{ij} stated without proof in (Cooke et al., 2011, Chapter 3) and identify the angles as the inverse cosine of the semi-partial correlations (SPCs) $\rho_{ji:1,2,...,j-1}$ between the variables y_i and y_j $(i > j)$ conditioned on the previous variables, see Huber (1981), Eaves & Chang (1992), and Madar (2015). Surprisingly, the simplicity of the relations between the angles and SPCs is reminiscent of the relations in (2.6) between the entries of the first columns of Θ and R.

Theorem 2.2.1. *Let R be a general $k \times k$ positive-definite correlation matrix with the Cholesky decomposition $R = BB^\top$ where the Cholesky factor B is a lower triangular matrix. Then,*
(a) the entries of $B = (b_{ij})$ can be expressed in terms of the semi-partial correlations (SPCs) as

$$b_{i1} = r_{i1}, \;\; b_{ii} = \sqrt{1 - \sum_{u=1}^{i-1} b_{iu}^2} \;\; \textit{for } i = 2, 3, ..., k, \tag{2.10}$$

$$b_{ij} = \rho_{ji:1,2,...,j-1} \prod_{u=1}^{j-1} \sqrt{1 - \rho_{ui:1,2,...,u-1}^2} \; \textit{for } 2 \le j < i \le k, \tag{2.11}$$

.

(b) the angles θ_{ij}'s are precisely the inverse cosine of the SPCs:

$$\rho_{ji:1,2,...,j-1} = cos(\theta_{ij}) \text{ for } 1 \leq j < i \leq k. \tag{2.12}$$

Proof: See Appendix A.

2.2.3 Distributions of the angles

Cholesky decomposition of a correlation matrix, and hence the concepts of the angles and semi-partial correlations depend on ordering or labeling the variables in R. Next, one may assign distributions to the angles so that the distribution of R is a power of its determinant and hence invariant to permutations of its rows and columns (Pourahmadi & Wang, 2015, Theorem 1).

Theorem 2.2.2. *For a k-dimensional random correlation matrix R with the corresponding matrix of angles Θ, let the random variables in the j^{th} column of Θ be independent and identically distributed as*

$$\theta_{ij} \sim p_j(\theta) \propto (sin\ \theta)^{2\alpha+k-j} \text{ for } j = 1, \cdots, k, \;\; i = j+1, j+2, ..., k, \tag{2.13}$$

where α is a constant, $\theta \in [0, \pi)$. Then

(a) the joint distribution of R is given by

$$p(R) = c_k(\alpha)[det(R)]^{\alpha}, \;\; c_k(\alpha) = \prod_{j=1}^{k-1} \Big(\frac{\Gamma(\frac{2\alpha+j}{2}+1)}{\sqrt{\pi}\Gamma(\frac{2\alpha+j+1}{2})} \Big)^j, \tag{2.14}$$

where $c_k(\alpha)$ is the normalizing constant.

(b) The marginal density of each r_{ij}, $1 \leq j < i \leq k$, of the correlation matrix R is proportional to $(1-r_{ij}^2)^{\alpha+\frac{k}{2}-1}$, i.e., a shifted Beta$(\alpha + k/2, \alpha + k/2)$ distribution in $[-1, 1]$.

(c) The distribution is symmetric about $\pi/2$, hence its mean and median are equal to $\pi/2$.

It turns out that these distributions on the angles reduce to the marginal uniform and joint

uniform priors of Barnard et al. (2000) on a correlation matrix for specific values of α. Recall that the *marginal uniform prior* assigns a shifted Beta distribution in $[-1, 1]$ to each r_{ij}:

$$p_M(r_{ij}) \propto (1 - r_{ij}^2)^{\nu-k-1}, |r_{ij}| \leq 1, \tag{2.15}$$

and *joint uniform prior* assigns a uniform distribution to the set of all valid $k \times k$ correlation matrices

$$p_J(R) \propto 1, R \in \mathcal{R}^k. \tag{2.16}$$

Indeed, $\alpha = 0$ in (2.14) leads to p_J, and $\alpha = \nu - 3k/2$ in (2.13) reduces to p_M. Note that the marginal uniform prior for each r_{ij} is peaked more at 0 for higher k. As such this prior is noninformative and not suitable in longitudinal data analysis, since higher lag (auto)correlations tend to zero faster than those with smaller lags.

2.3 Bayesian estimation of a Correlation Matrix

We assume throughout that the data $Y_1, Y_2, ..., Y_n$ follow a normal distribution $N(0, R)$. Restricting attention to correlation matrices is natural, for example, in the analysis of multivariate probit model to circumvent the issue of identifiability (Chib & Greenberg, 1998). The key intuition behind our prior elicitation for correlation matrix of longitudinal data is that one expects two variables far apart have correlation decaying to zero. Therefore, it is natural to expect that the semi-partial correlation between two variables y_i & y_j $(i > j)$ in the random vector Y given the preceding variables $y_1, y_2, ..., y_{j-1}$ decays to zero as the lag (i-j) increases. In terms of angles, this essentially means that the corresponding θ_{ij} goes to $\pi/2$, since θ_{ij} is related to the corresponding semi-partial correlation only through the cosine function (2.12). This simple observation serves as the main guide for various priors for θ_{ij}'s. In this section, we work with four priors and study their properties and numerical performances in estimating a correlation matrix R. For comparison, we focus on the angle counterparts of shrinkage and selection priors on PAC in Gaskins et al. (2014). A key role is played by the (modified) shifted beta distribution, denoted by SBeta, in $[-1, 1]$:

$$p(y) = \frac{(1+y)^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)2^{\alpha+\beta-1}} \quad , \qquad \text{for } -1 \leq y \leq 1, \tag{2.17}$$

with parameters α, β.

We note that in Bayesian statistics, spike and slab priors have also been used in practice as a selection prior with a spike at a target value, say $\pi/2$. There is a vast literature on selection priors (Mitchell & Beauchamp, 1988; Ishwaran & Rao, 2005).

2.3.1 Selection and shrinkage priors for the angles

Next, we propose selection and shrinkage priors for the angles. Our prior elicitation is motivated by noting that in longitudinal data two variables which are far apart have correlation decaying to zero. Therefore, it is natural to expect that the semi-partial correlation between two variables y_i & y_j with $i > j$ given the preceding variables $y_1, y_2, ..., y_{j-1}$ decays to 0 as the lag (i-j) increases. For the angles, this essentially means that the corresponding θ_{ij} is expected to be around $\pi/2$, see the identity (2.12).

2.3.1.1 Selection priors

A way to motivate the formation of our selection prior, note that when $R = I_k$ an identity matrix of order k, then all the entries of Θ are $\pi/2$. Thus, forming a selection prior as a mixture of a Dirac delta with mass at $\pi/2$ and a continuous density having support in $[0, \pi)$, is capable of selecting or centering the angles at $\pi/2$. In terms of the SPCs, this amounts to encouraging the semi-partial correlation between y_i and y_j given $y_1, y_2, ..., y_{j-1}$ to be centered at 0.

We recall that in the PAC framework, a selection prior in Gaskins et al. (2014) for π_{ij} is:

$$p_{\pi,S}(\pi_{ij}; \alpha, \beta) \sim \eta_{ij}\text{Sbeta}(\alpha, \beta) + (1 - \eta_{ij})\delta_0 \tag{2.18}$$

where δ_0 is the Dirac delta function with mass at 0 and Sbeta is a shifted beta distribution. Our selection prior on the angles, denoted by $p_{\theta,S}$, assumes independent mixture distributions for indi-

vidual θ_{ij}'s by

$$p_{\theta,S}(\theta) \propto (1-\eta_{ij})\delta_{\pi/2}(\theta) + \eta_{ij}(\sin(\theta))^{k-j}, \text{ where } \theta \in [0,\pi] \tag{2.19}$$

where, $\eta_{ij} = Pr(\theta_{ij} \neq 0)$ for $1 \leq i < j \leq k$ and $\delta_{\pi/2}$ denotes a Dirac delta with mass at $\pi/2$. To make such priors more suitable for longitudinal data, we further parameterize $\eta_{ij} = \eta_0|j-i|^{-\gamma}$ so that as the lag $|j-i|$ increases, the prior in (2.19) puts more weight at $\pi/2$. Since the angle θ_{ij} is related to the partial correlation $\rho_{ji:1,2,\ldots,j-1}$ in (7) through $\cos(\theta_{ij})$, this implies that for variables which are far apart or having greater lag $|i-j|$, the corresponding θ_{ij}'s are closer to $\pi/2$. We further assume a Unif$(0,1)$ distribution for the hyper-parameter η_0 and a Gamma(a,a) distribution for the hyper-parameter γ so that γ has prior mean 1. For the continuous component in (2.19), one may use a wrapped exponential distribution restricted to $[0,\pi)$, instead of a multiple of $(\sin(\theta))^{k-j}$. In our simulation study, we choose $a = 5$ to make our results comparable to those in Gaskins et al. (2014).

2.3.1.2 Shrinkage priors

In Bayesian covariance estimation, shrinkage priors have been used to shrink the posterior estimate towards specific structures. For example, Liechty et al. (2004) considered priors to shrink the correlation matrix to certain group-structured targets, and Wang & Pillai (2013) considered scale mixture of uniform distributions to construct shrinkage priors for covariance matrix estimation.

The shrinkage prior in Gaskins et al. (2014) shrinks the PAC (π_{ij})'s towards 0 using a shifted beta density in $[-1,1]$, namely, $\pi_{ij} \sim$ Sbeta(α_{ij},β_{ij}). When $\alpha_{ij} = \beta_{ij}$, then $E(\pi_{ij}) = 0$, and its variance for general parameter values is given by

$$Var(\pi_{ij}) = \frac{4\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij})^2(\alpha_{ij}+\beta_{ij}+1)} = \xi_{ij}. \tag{2.20}$$

In the interest of parsimony, they parameterize $\xi_{ij} = \xi_0|i-j|^{-\gamma}$, for $\xi_0 \in (0,1), \gamma > 0$, so that for

longitudinal data higher-lag terms are shrunk to 0 more aggressively.

Defining an analogue of the above shrinkage prior on angles using Sbeta distribution runs into difficulty as computing the mean and variance does not seem easy. For the time being we resort to some distributions from directional statistics.

2.3.1.3 Shrinkage priors from directional statistics

Since the support of θ_{ij} is $[0, \pi)$, one may use the reservoir of distributions from directional statistics (Mardia & Jupp, 2009) as possible priors.

For example, the truncated wrapped exponential distribution with a nonegative parameter λ_{ij}:

$$\theta_{ij} \sim p(\theta) = \frac{\lambda_{ij}\text{exp}(-\lambda_{ij}\theta)}{1 - \text{exp}(-\pi\lambda_{ij})}, 0 < \theta < \pi, \tag{2.21}$$

is a plausible prior for θ_{ij}. Its mean is given by $E(\theta_{ij}) = \arctan(1/\lambda_{ij})$, where we further parameterize $\lambda_{ij} = \lambda_0|j - i|^{-\gamma}$, where $\lambda_0, \gamma > 0$. A distinctive feature of this parametrization and the proposed prior is that as the lag $|j - i|$ increases, λ_{ij} gets smaller and thus the prior mean $E(\theta_{ij}) = \tan^{-1}(1/\lambda_{ij})$ approaches to $\pi/2$. This is consistent with the fact that semi-partial correlation between y_j and y_i given $y_1, y_2, ..., y_{i-1}$ approaches to zero for higher lags. For the hyper-parameters λ_0, γ, we assume $\lambda_0 \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Gamma}(a, a)$ distribution so that γ has prior mean 1 and thus λ_0 has a subtle role in determining the prior mean which is a decreasing function of λ_{ij}. We note that higher lags play the major role in determining λ_{ij} and the influence of λ_0 gradually becomes prominent on λ_{ij} as the lag decreases. For simulation study, we choose $a = 5$ to compare our results with those by Gaskins et al. (2014). Otherwise, one can assume a further level of uncertainty by using a hyper-prior distribution on a. As an alternative prior one may take the von Mises distribution in Mardia & Jupp (2009)

.

2.3.2 Posterior computation using Metropolis-Hastings scheme

For sample data $y_1, y_2, ..., y_n$ coming from a k dimensional normal distribution with mean 0 and covariance matrix R, the likelihood function parameterized by the angles Θ (using the transformation T_b) is given by

$$L(y_1, y_2, ..., y_n|\Theta) \propto \det(T_b(\Theta))^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n} y_i (T_b(\Theta))^{-1} y_i\}. \tag{2.22}$$

Here, Θ is the parameter of interest and we denote the hyper-parameters by a generic vector $\boldsymbol{\psi}$. For the prior models 2.15 and (2.16), $\boldsymbol{\psi}$ is of course empty since we have no further hyper-parameters. For the shrinkage prior given by (2.21), $\boldsymbol{\psi} = (\lambda_0, \gamma)$; for the selection prior given by (2.19), $\boldsymbol{\psi} = (\eta_0, \gamma)$. The posterior distribution is then

$$p(\Theta, \boldsymbol{\psi}|y_1, y_2, ..., y_n) \propto L(y_1, y_2, ..., y_n|\Theta) p(\Theta|\boldsymbol{\psi}) p(\boldsymbol{\psi}) \tag{2.23}$$

where the forms of $p(\Theta|\boldsymbol{\psi})$ and $p(\boldsymbol{\psi})$ can be specified through the priors. We note that θ_{ij}'s (entries of Θ) appear nonlinearly (are badly entangled) in the posterior distribution, and hence there is no closed form for the conditional distribution of θ_{ij}, $p(\theta_{ij}|\boldsymbol{\psi}, \Theta_{[-i,j]}, y_1, y_2, ..., y_n)$. We perform a Metropolis-Hastings scheme to obtain posterior estimate of Θ, and rely on the following component-wise Metropolis-Hastings (Roberts & Rosenthal, 2009) to update a single θ_{ij} and $\boldsymbol{\psi}$ at a time:

Metropolis-Hastings scheme:

(1) Initialization: Start with initial values $\Theta^{(0)}, \boldsymbol{\psi}^{(0)}$.

(2) Proposal distributions:

I. Sample $\theta_{i,j}^{(cand)}$ from a Unif$(0, \pi)$ distribution. Construct a lower triangular matrix $\Theta^{(cand)}$ equals to Θ^0 except (i, j)-th entry is replaced by $\theta_{i,j}^{(cand)}$.

II. Sample $\boldsymbol{\psi}^{(cand)}$ from proposal density $q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(0)})$ which is the product of $q(\eta_0|\eta_0^{(0)})$ and $q(\gamma|\gamma^{(0)})$ for selection prior and product of $q(\lambda_0|\lambda^{(0)})$ and $q(\gamma|\gamma^{(0)})$ for shrinkage prior.
For $q(\eta_0|\eta_0^{(0)})$ and $q(\lambda_0|\lambda_0^{(0)})$, we choose $Unif(a,b)$, where $a = \max\{0, \eta^{(0)}(\text{or } \lambda^{(0)}) - 0.05\}$ and $b = \min\{\eta^{(0)}(\text{or } \lambda^{(0)}) + 0.05, 1\}$ and for $q(\gamma|\gamma_0)$, we choose a Gamma(5,5) distribution.

(3) Jump from $(\Theta^0, \boldsymbol{\psi}^0)$ to $(\Theta^{(cand)}, \boldsymbol{\psi}^{(cand)})$ with probability

$$\alpha = \min\Big\{1, \frac{p(\Theta^{(cand)}, \boldsymbol{\psi}^{(cand)}|y_1, y_2, ..., y_n)}{p(\Theta^0, \boldsymbol{\psi}^0|y_1, y_2, ..., y_n)} \frac{q(\boldsymbol{\psi}^{(0)}|\boldsymbol{\psi}^{(cand)})}{q(\boldsymbol{\psi}^{(cand)}|\boldsymbol{\psi}^{(0)})}\Big\}$$

(4) Repeat the steps (1-3) for other entries of Θ.

(5) Update $R = T_b(\Theta)$ for each iteration.

.

2.4 Simulations

We perform a number of simulation studies to assess the performance of our selection and shrinkage priors on the angles relative to the selection and shrinkage priors of Gaskins et al. (2014) on partial autocorrelations.

The frequentist risks of the posterior estimates are evaluated by averaging the loss over 60 simulation runs for the following two loss functions: The Kullback-Liebler loss function $L_1(R, \hat{R}) = \text{tr}(\hat{R}^{-1}R) - \log|\hat{R}^{-1}R| - k$, which is zero when $\hat{R} = R$. The second loss function for estimating Θ is defined by $L_2(\Theta, \hat{\Theta}) = ||\hat{\Theta} - \Theta||_F = \sum_{i<j}(\theta_{ij} - \hat{\theta}_{ij})^2$, where $\hat{R}$ and $\hat{\Theta}$ denote the posterior estimates.

2.4.1 Comparing priors on the angles and PACs

In this section we compare the performance of our priors on the angles with the shrinkage and the selection priors on PACs in Gaskins et al. (2014). Since the selection prior performed better than the shrinkage prior in their simulation study, here we focus only on the shrinkage prior and follow their simulation set-up as much as possible.

We consider 4 different 6×6 correlation matrices: I_6 (with the matrix Θ a lower-triangular matrix of order 5 having all entries $\pi/2$), AR(1) matrix with correlation 0.7, and R_C, R_D constructed from the following Θ matrices:

$$\Theta_C = \begin{bmatrix} \pi/4(0.707) & 0.866 & 0 & 0 & 0 \\ \pi/6 & \pi/4(0.862) & 0.612 & 0 & 0 \\ \pi/2 & \pi/6 & \pi/4(0.431) & 0.306 & 0 \\ \pi/2 & \pi/2 & \pi/6 & \pi/4(0.431) & 0.306 \\ \pi/2 & \pi/2 & \pi/2 & \pi/6 & \pi/6(0.459) \end{bmatrix},$$

$$\Theta_D = \begin{bmatrix} \pi/4(0.707) & 0.866 & 0.707 & 0.866 & 0.809 \\ \pi/6 & \pi/4(0.862) & 0.933 & 0.789 & 0.866 \\ \pi/4 & \pi/6 & \pi/2(0.829) & 0.838 & 0.848 \\ \pi/6 & \pi/3 & \pi/2 & \pi/2(0.765) & 0.827 \\ \pi/5 & \pi/4 & \pi/2 & \pi/2 & \pi/2(0.805) \end{bmatrix}.$$

It can be seen that Θ_C leads to a banded correlation matrix and the entries in the rows of Θ_D decay to $\pi/2$.

For each of the 4 correlation matrices, we simulate 60 samples from a normal distribution having mean zero and covariance matrix equals to the chosen correlation matrix. For comparison of the risks, our competitor is $p_{\pi;SP}$ which performed the best in Gaskins et al. (2014). We run an MCMC chain for 2000 iterations with a burn-in 500 and the posterior estimate of the correlations is obtained by taking the mean of the samples after burn-in. For each case, we gauged the performance by the risk estimates with respect to the two loss functions discussed earlier by taking average of these loss functions over 60 replications of the simulated data.

The results are summarized in Table 2.1 (results for $p_{\pi;SP}$ have been reprinted using the codes available in Gaskins et al. (2014)), where we note that for the identity matrix our selection prior outperforms all its competitors, but our shrinkage prior is outperformed by the selection prior in Gaskins et al. (2014). For the AR(1), our selection prior and the selection prior of Gaskins et al.

Table 2.1: Risks for our selection prior ($p_{\theta;SP}$), shrinkage prior ($p_{\theta;SH}$) and the selection prior in Gaskins et al. (2014).

n	R	Loss	$p_{\pi;SP}$	$p_{\theta;SH}$	$p_{\theta;SP}$
20	I_k	$L_1(\hat{R}, R)$	0.025	0.081	0.023
		$L_2(\hat{\Theta}, \Theta)$	0.6027	0.8697	0.0802
200	I_k	$L_1(\hat{R}, R)$	0.0014	0.0077	0.0011
		$L_2(\hat{\Theta}, \Theta)$	0.0422	0.2700	0.0325
20	AR(1)	$L_1(\hat{R}, R)$	0.34	0.53	0.064
		$L_2(\hat{\Theta}, \Theta)$	0.6113	0.5998	0.2581
200	AR(1)	$L_1(\hat{R}, R)$	0.027	0.057	0.054
		$L_2(\hat{\Theta}, \Theta)$	0.1501	0.1444	0.0794
20	R_C	$L_1(\hat{R}, R)$	2.095	0.6408	0.5749
		$L_2(\hat{\Theta}, \Theta)$	0.9532	0.3527	0.2840
200	R_C	$L_1(\hat{R}, R)$	1.665	0.0947	0.0466
		$L_2(\hat{\Theta}, \Theta)$	0.7562	0.0788	0.0682
20	R_D	$L_1(\hat{R}, R)$	2.0035	0.8848	0.5749
		$L_2(\hat{\Theta}, \Theta)$	0.9779	0.6966	0.3558
200	R_D	$L_1(\hat{R}, R)$	1.5001	0.0782	0.0458
		$L_2(\hat{\Theta}, \Theta)$	0.5672	0.2132	0.0957

(2014) are comparable. For banded R_C, our selection prior and shrinkage prior are comparable and perform better than $p_{\pi;SP}$. Finally, for R_D our selection prior outperforms $p_{\pi;SP}$ and our shrinkage prior.

In summary, our selection prior and shrinkage prior show advantage over based on certain scenarios.

2.4.2 Comparing priors on the angles with p_M and p_J.

We assess the performance of our priors on the angles relative to the marginal uniform prior (p_M) and joint uniform prior (p_J) in Barnard et al. (2000). For the marginal uniform prior obtained through (2.13), we choose $\alpha = 0.1$.

We consider three different settings for the true correlation matrix R, namely the identity matrix, the AR(1) correlation matrix of the form $r_{ij} = 0.4^{|i-j|}$, and a general correlation matrix R_k of dimension k. A general random correlation matrix is generated using the method in Pourahmadi & Wang (2015).

For each of these correlation matrices, we consider three settings for (n, k), namely $(100, 5)$, $(500, 10)$ and $(1000, 15)$ where n and k denote the sample size and dimension of the correlation matrix. We simulated n samples from a k-variate zero-mean normal distribution with the covariance matrix equal to the chosen R. Having expressed the likelihood in terms of Θ (2.22) and using prior, we calculate the posterior of Θ along with the set of hyper-parameters $\boldsymbol{\psi}$ according to (2.23). For each data-set, an MCMC chain is run with 2000 iterations with a burn-in of 500 following the Metropolis-Hastings scheme with the posterior in (2.23). For this comparison, we replicated the MCMC chain 50 times and took the average loss over those replications. The results summarized in Table 2.2 show that our proposed selection prior outperforms the marginal uniform prior and joint uniform prior of Barnard et al. (2000) in terms of both risks.

The performance of our selection prior is remarkable for the identity matrix in all dimensions. This is reasonable since selection prior is capable of selecting 0's which are the essentially all the entries of identity matrix. For the AR(1) matrix, the performance although is not as remarkable as that of the identity matrix, it indeed supersedes the other two priors. For $k = 5$, the performance is

Table 2.2: Risks of the marginal uniform prior (p_M), joint uniform prior (p_J), selection prior ($p_{\theta;SP}$)

(n,k)	R	Loss	p_M	p_J	$p_{\theta;SP}$
$(50,5)$	I_k	$L_1(\hat{R},R)$	0.1688	0.1171	0.0019
		$L_2(\hat{\Theta},\Theta)$	0.4095	0.3413	0.0445
$(100,10)$	I_k	$L_1(\hat{R},R)$	0.4885	0.3198	0.0051
		$L_2(\hat{\Theta},\Theta)$	0.6963	0.5647	0.0712
$(500,15)$	I_k	$L_1(\hat{R},R)$	0.5582	0.4950	0.0073
		$L_2(\hat{\Theta},\Theta)$	0.5976	0.6081	0.0854
$(50,5)$	AR(1)	$L_1(\hat{R},R)$	0.2048	0.2750	0.1048
		$L_2(\hat{\Theta},\Theta)$	0.3988	0.4130	0.3233
$(100,10)$	AR(1)	$L_1(\hat{R},R)$	0.2695	0.2778	0.2015
		$L_2(\hat{\Theta},\Theta)$	0.3781	0.4303	0.2125
$(500,15)$	AR(1)	$L_1(\hat{R},R)$	0.2709	0.2838	0.2083
		$L_2(\hat{\Theta},\Theta)$	0.4762	0.4945	0.4000
$(50,5)$	R_5	$L_1(\hat{R},R)$	0.2041	0.1833	0.1788
		$L_2(\hat{\Theta},\Theta)$	0.4233	0.5179	0.3774
$(100,10)$	R_{10}	$L_1(\hat{R},R)$	0.5373	0.4551	0.3223
		$L_2(\hat{\Theta},\Theta)$	0.6052	0.5535	0.4983
$(500,15)$	R_{15}	$L_1(\hat{R},R)$	1.4238	1.3814	0.8275
		$L_2(\hat{\Theta},\Theta)$	0.7725	0.7645	0.5000

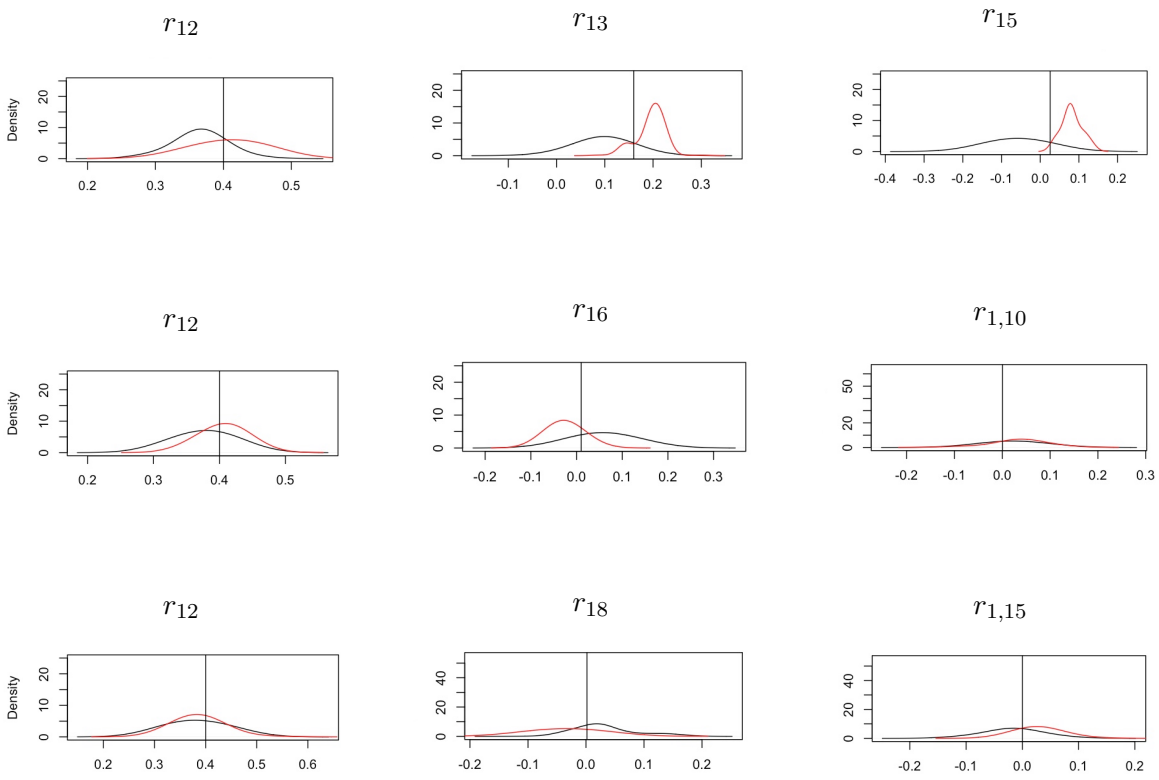

Figure 2.1: Posterior density plots of indicated entries of the AR(1) correlation matrix. The rows correspond to dimensions $k = 5, 10, 15$, the black curves pertain to marginal uniform prior, red to selection prior and vertical lines correspond to true values.

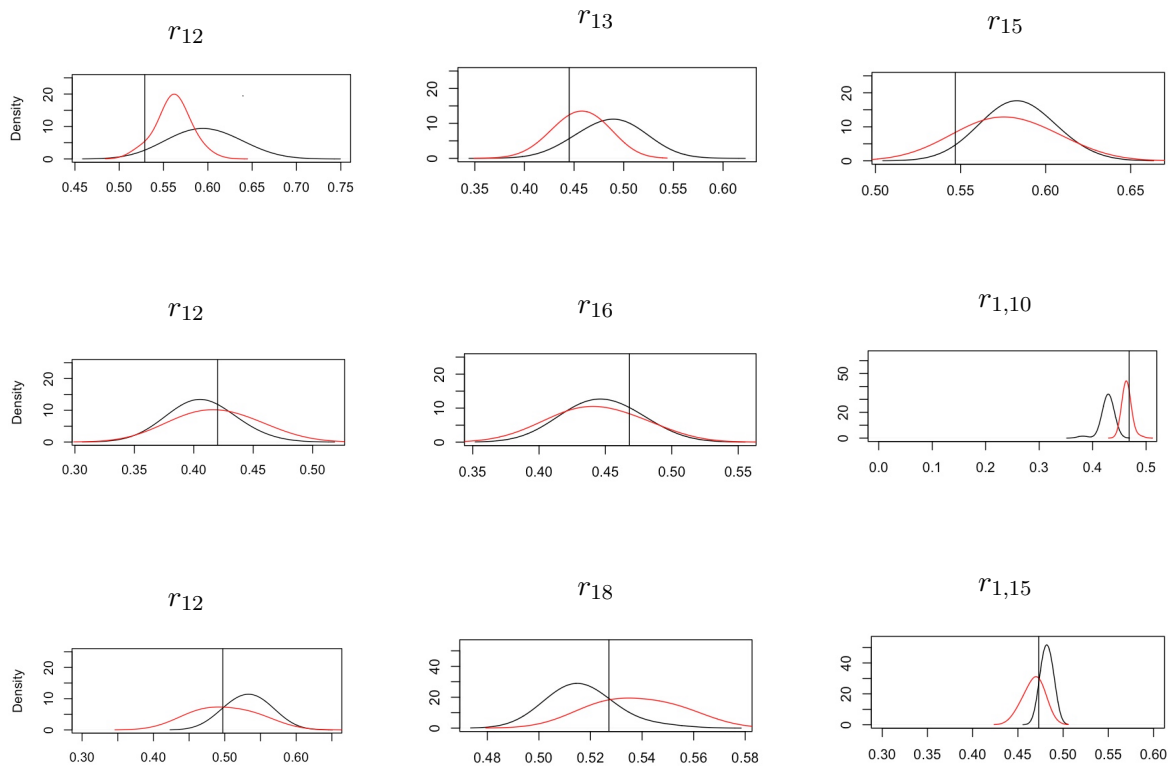

Figure 2.2: Posterior density plots of indicated entries of R_5, R_{10}, R_{15}. The rows correspond to dimensions $k = 5, 10, 15$, the black curve pertains to marginal uniform prior, red to the selection prior and vertical line corresponds to true values.

50% better and for $k = 10, 15$ the performance is 25% better than Barnard's priors with respect to the loss function L_1.

For a general correlation matrix R_k with $k \in \{5, 10, 15\}$, our selection prior clearly outperforms marginal uniform prior and performs as good as joint uniform prior with respect to the loss function L_1. The main reason for the good performance of the joint uniform prior is that it is essentially a non-informative prior since the only hyper-parameter α is set to 0. Thus, the posterior is solely influenced by the data. On the other hand, using an informative prior (selection prior) we get performance as good as the joint uniform prior.

We plot the posterior density of the indicated elements obtained from the MCMC for AR(1) and general correlation matrices of different dimensions in Figure 2.1 and 2.2 respectively. The vertical line in these plots indicate the true value of the element. It is evident that the posterior density plots arising out of the selection prior concentrate more around the true value of the elements than the marginal uniform prior.

2.4.3 Computational advantages of angle parameterization

The computational challenges of using constrained priors like the joint uniform prior $p_J(R)$ are well-known, other notable examples are the common correlation priors in Liechty et al. (2004), priors for sparse R^{-1} in Wong et al., 2003; Pitt et al., 2006; Carter et al., 2011, which place a flat prior on the non-zero components for a given pattern of zeros. These methods usually require computing the normalizing constants related to volumes of certain subsets of $\mathcal{R}^k$ corresponding to patterns of zeros, and where the prior and posterior densities are supported on constrained sets. Due to the presence of the indicator function of $\mathcal{R}^k$ in the prior and posterior, in the Metropolis-Hastings scheme, the proposal density for updating r_{ij} has to be restricted to an interval $[l_{ij}, u_{ij}]$ where these bounds are functions of the rest of the entries of $R^{\pm 1}$ (Barnard et al. (2000), Liechty et al. (2004)). Of course, unconstrained parameterization resolves the tedious task of computing the normalizing constant in every update of the MCMC algorithm and consequently posterior computation is faster.

Next, we compare the time complexity of implementing the MCMC algorithm for the constrained prior $p_J(R)$ on the space of valid correlation matrices $\mathcal{R}^k$, and its two unconstrained

reparameterizations on the spaces of angles Θ and partial autocorrelations $\mathbf{\Pi}$. The prior on the angles:

$$p_{\theta J}(\Theta) \propto \prod_{i>j} (\sin \theta_{ij})^{k-j}, \tag{2.24}$$

is obtained from (2.13) for $\alpha = 0$, and the prior on PACs is

$$p_{\pi J}(\mathbf{\Pi}) \propto \prod_{i<j} (1-\pi_{ij}^2)^{1+[(k-1)-(j-i)]/2}, \tag{2.25}$$

where $\pi_{ij} \in [-1, 1]$, for more details see Gaskins et al. (2014).

We consider three different settings of (n, k), namely $(50, 5)$, $(100, 10)$ and $(500, 15)$ and simulate a sample of size n from a k-dimensional normal distribution having mean 0 and covariance matrices set to Identity, AR(1) with correlation 0.4 and a general correlation matrix, respectively. In Figure 2.3, we present run times (in log scale) for 1000 iterations of MCMC for computing the posterior of R. As expected the unconstrained priors outperform constrained method significantly in any dimension with respect to the execution-time. The simulations were run on a 2.6 GHz Intel Core i5 processor. The numerical results above may not be surprising by noting that the computational complexity of simulating a posterior of R based on priors on angles or generating general random correlation matrices (Pourahmadi & Wang, 2015) is $O(k^3)$ compared to $O(k^4)$ of the Joe (2006) proposal based on partial correlations, and $O(k^3)$ of the Lewandowski et al. (2009) method using the partial correlations defined on C-vines, respectively.

2.5 Data Analysis

We analyze a data set (Gaskins et al., 2014) simulated based on first Commit to Quit (CTQ I) study of Marcus et al. (1999), a clinical trial designed to encourage women to stop smoking. The aim of the study was how exercise is effective to increase quit rate, as weight gain seems to be an influencing factor in a smoking cessation program. Providing an educational intervention of equal time for the control group, the study spans 12 weeks and the patients were encouraged to quit smoking at week 5.

The data is provided in the form of a 281×9 matrix, where rows correspond to patients and

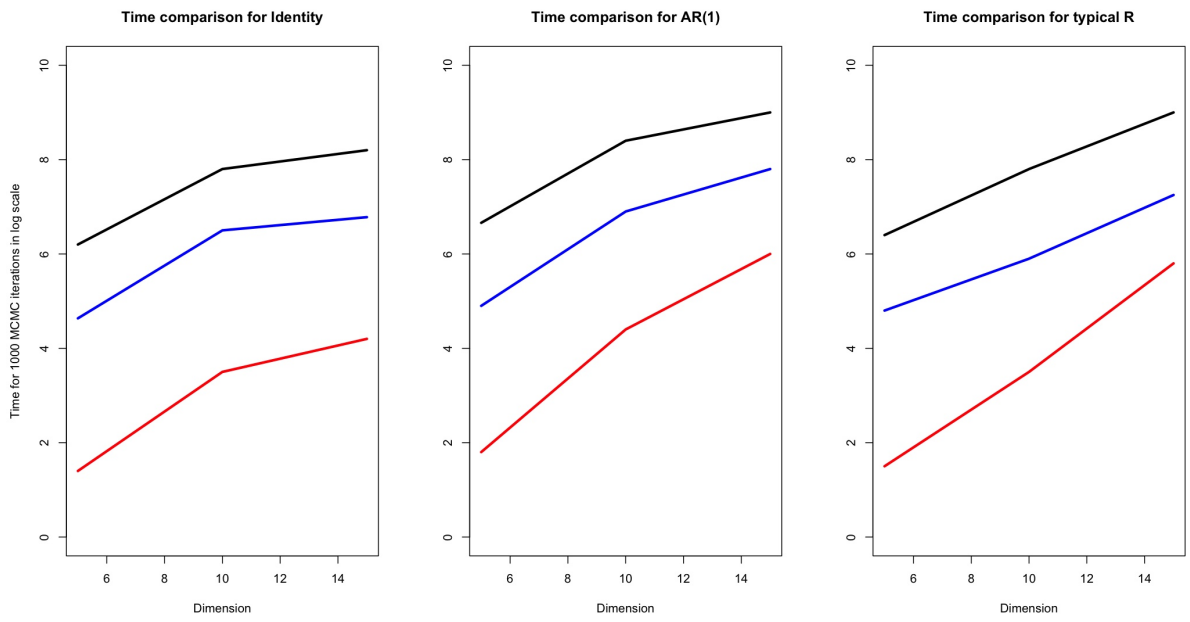

Figure 2.3: Time comparison in log scale for constrained vs unconstrained method for 1000 iterations of MCMC algorithm for three indicated correlation matrices. The black line indicates constrained prior p_J, blue line indicates unconstrained prior for $\mathbf{\Pi}$ and red line for unconstrained prior on Θ.

columns 2-9 correspond to weeks and first column corresponds to treatment assignment (0 for control and 1 for exercise). For each patient, columns 2-9 denote the patient's smoking status from 5-th to 12-th week after they are asked to quit smoking. With $n = 281, k = 8$ (discarding first column), we associate an $n \times k$ matrix $Y = (y_{ij})$ to the data, whose entries take values -1,0,1; where 1 denotes success (i-th patient not smoking in j-th week), -1 denotes failure (still smoking in j-th week) and 0 denotes a missing observation. Introducing latent variables y^*_{ij}, we assume a multivariate probit model Chib & Greenberg (1998) where,

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} > 0, \\ -1 & \text{if } y^*_{ij} < 0, \end{cases}$$

and if $y_{ij} = 0$, the sign of y^*_{ij} represents the (unobserved) quit status for the week.

Next, we assume $y_i^* = (y_{i1}^*, y_{i2}^*, \cdots, y_{ik}^*)^\top \sim N_k(\mu_i, R)$ for $i = 1, 2, \cdots, n$ and μ_i is paramterized as $\mu_i = X_i^\top \beta$; where X_i is a $q \times k$ matrix of covariates and β is a $q \times 1$ vector of regressors. To circumvent identifiability issue, covariance matrix is restricted to be a correlation matrix. As in Gaskins et al. (2014), we consider two choices of X_i: time-varying which specifies a different μ_{it} for each time within each treatment group $(q = 2k)$ and time-constant which gives the same μ_{it} across all times within treatment group $(q = 2)$. With this set-up, we consider a flat prior on β and the priors on R are the selection and shrinkage priors in Gaskins et al. (2014) for PACs and the angle (Θ), respectively.

2.5.1 Posterior Computation

For posterior computation, we run an MCMC chain for 12,000 iterations with a burn-in of 3000, retaining every tenth observation. The three sets of parameters appeaing in the posterior are regression parameters, latent variables and correlation matrix.

1. Sampling β. The conditional posterior of β given latent variables $y_i^*, i = 1, 2, ..., n$ and R is multivariate normal.

2. Sampling R. For angle based priors, Metropolis-Hastings scheme in 2.3.2 is used and R code provided in Gaskins et al. (2014) has been used for PACF based priors on the residuals $y_i^* - \mu_i$, for $i = 1, 2, \cdots, n$.

3. Sampling y_i^*s. For sampling latent variables, we use Proposition 1 of Liu et al. (2009) as in Gaskins et al. (2014).

For comparison we use deviance information criterion (DIC) which does not require counting the number of model parameters, making it an effective criterion for model selection when shrink-

age or sparsity is concerned. DIC is defined as (Spiegelhalter et al., 2002) Dev $+ 2p_D$ where,

$$\text{Dev} = -2\sum_{i=1}^{n} l(\hat{\beta}, \hat{R}|y_i) \tag{2.26}$$

$$p_D = E\{-2l(\beta, R|Y)\} - \text{Dev} \tag{2.27}$$

with l denoting log-likelihood function and expectation is taken with respect to the posterior distribution.

For the CTQ data, the posterior estimate $\hat{\beta}$ is the posterior mean, as for the posterior estimate of $\hat{R}$ we use the posterior median for angle-based priors and the one used by (Gaskins et al., 2014, pp.12) for PAC-based priors. The numerical results for various priors on the correlation matrix are reported in Table 2.3, where it can be seen that the DIC is smaller for the time constant mean structure in coherence with the findings of Gaskins et al. (2014). One can note that for time varying mean structure, the models are heavily penalized by p_D which deals with 14 extra parameters compared to time constant models. Among the priors, our angle-based selection prior appears to be the best with the DIC value of 1052, the lowest in Table 2.3.

Table 2.3: DICs for various correlation priors for CTQ data

Mean Structure	Prior	Dev	p_D	DIC
Time Constant	$p_{\pi;SH}$	1027	14	1058
Time Constant	$p_{\pi;SP}$	1045	13	1070
Time Constant	$p_{\theta;SH}$	1022	13	1057
Time Constant	$p_{\theta;SP}$	1030	11	1052
Time Varying	$p_{\pi;SH}$	1017	25	1069
Time Varying	$p_{\pi;SP}$	1037	29	1075
Time Varying	$p_{\theta;SH}$	1030	23	1063
Time Varying	$p_{\theta;SP}$	1015	20	1057

2.6 Discussion

We have dealt with some computational challenges in Bayesian estimation of correlation matrices by using its Cholesky decomposition and the ensuing angles as the new parameters which vary freely in $[0, \pi)$. This reparametrization deals effectively with the positive-definiteness constraint on a correlation matrix and results in faster computation of the posteriors. At a first encounter, angles may not seem the most natural parameters in statistics. However, to our knowledge we have shown for the first time that the angles in the present context are simply the inverse cosine of the familiar semi-partial correlations, see Huber (1981), Eaves & Chang (1992), Cooke et al. (2011). Thus, the angles are statistically meaningful and the new connection opens up the possibility of using the wealth of distributions from directional statistics as potential priors for Bayesian analysis of correlation matrices. Through simulations and data analysis we have shown that the performance of our shrinkage and selection priors on the angles is better or comparable to those based on the PACs in Gaskins et al. (2014) and marginal and joint priors in Barnard et al. (2000).

3. CHARACTERIZATION OF STRUCTURED CORRELATION MATRICES AND BAYESIAN VARIABLE CLUSTERING

In this chapter, we first characterize some structured correlation matrices through structured angular matrices and later exploit that to cluster variables based on block diagonal correlation with equicorrelated blocks.

3.1 Characterization of Structured Correlation Matrices

3.1.1 Compound Symmetric structure

A k-dimensional correlation matrix has compound symmetric structure if all of its off-diagonal entries equal to a common value r where $-1/(k-1) < r < 1$. The corresponding angular matrix Θ is characterized by a single angle θ and the remaining angles can be expressed explicitly as a function of θ. The relationship between θ and r is precisely $r = \cos(\theta)$.

3.1.2 AR(1) structure

A k-dimensional AR(1) matrix R is of the form $r_{ij} = r^{|i-j|}$, for $1 \leq i, j \leq k$, where $-1 < r < 1$. The corresponding angular matrix Θ here too is characterized by a single angle θ, where $r = \cos(\theta)$. We can repeat the same inductive argument to verify this.

3.1.3 Hub Correlation model

Hardin et al. (2013) considered a Hub observation model-based on a single hub-observation and the relationship of each observation to that original hub. Each observation in a group is correlated with the hub-observation in a decreasing manner. We let the first variable corresponds to the hub-observation and consider a single group in which remaining variables belong to. Thus, we need to compute the first row and hence the first column r_{i1} for $j = 2, 3, ..., k$ according to the Hub structure and remaining entries of R will be determined such that it will be positive definite.

In particular, one may assume $r_{11} = 1$ and $r_{j1} = \rho_{max} - (\rho_{max} - \rho_{min})\left(\frac{j-2}{k-2}\right)^{\gamma}$ for $j = 2, 3, ..., k$ so that $r_{21} = \rho_{max}$ and $r_{k1} = \rho_{min}$ where ρ_{max} and ρ_{min} denote the maximum and minimum

correlation value between the hub and other observations. Clearly, r_{j1} decays and decay rate depends on γ, e.g. $\gamma = 1$ implies that the decay rate is linear.

The Hub correlation matrix can be characterized by the angular matrix Θ where the first row of Θ needs to be computed according to the Hub-correlation structure and remaining entries of Θ can be set to $\pi/2$. The angular matrix Θ can be identified by three (two) pivotal angles according to the case $\gamma \neq 1 (\gamma = 1)$. To see this, we recourse to the Cholesky decomposition $R = BB^\top$, $r_{j1} = b_{11}b_{j1} = b_{j1} = \cos(\theta_{j1})$ for $j = 2, 3, ..., k$. Thus two obvious pivotal angles will be $\theta_{21} = \arccos(\rho_{max})$ and $\theta_{k1} = \arccos(\rho_{min})$. For $\gamma \neq 1$, we need one additional pivotal angle $\theta_{31} = \arccos(\rho_{max} - \rho_{min}\left(\frac{1}{k-2}\right)^{\gamma})$. For the remaining entries of R, $r_{ij} = \sum_{l=1}^{j} b_{il}b_{jl} = b_{i1}b_{j1} = \cos(\theta_{i1})\cos(\theta_{j1})$, since $\theta_{ij} = \pi/2$ for $i \neq 1, i \neq j$. The resulting matrix R is of course a positive definite matrix thanks to this unconstrained parametrization.

However, Hardin et al. (2013) considered a Toeplitz or AR(1) structure to fill up rest of the matrix. Specifically they considered the following partitioned structure

$$R = \begin{bmatrix} 1 & r_1^\top \\ r_1 & \tilde{R} \end{bmatrix}$$

where $r_1 = (r_{12}, r_{13}, ..., r_{1k})^\top$ and $\tilde{R} = ((\tilde{r}_{ij}))$ is the $(k-1) \times (k-1)$ correlation matrix corresponding to $2, 3, ..., k$-th variables and $\tilde{r}_{ij} = \rho^{|i-j|}$. Since $\tilde{R}$ has a AR(1) structure, one needs one additional angle to characterize $\tilde{R}$ following the discussion on AR(1) structure. Therefore, one needs four (three) pivotal angles to characterize R according to $\gamma \neq 1 (\gamma = 1)$.

3.1.4 Banded Correlation matrix

In this section, we will characterize a banded correlation matrix.

Definition 3.1.1. *We say a k dimensional correlation matrix $R = ((r_{ij}))$ is λ-banded for $1 \leq \lambda < k$ if R is of the form*

$$r_{ij} = \begin{cases} r_{ij} \neq 0, & \textit{if } |i-j| \leq \lambda \\ 0, & \textit{otherwise} \end{cases} \tag{3.1}$$

We give a proposition below which connects a banded correlation matrix with the corresponding angular matrix.

Proposition 3.1.1. *A k dimensional correlation matrix R is λ-banded if and only if the corresponding angular matrix Θ as specified in Section 2 satisfies $\theta_{ij} = \pi/2$ for $|i-j| > \lambda$.*

The proof of the proposition is deferred to appendix A.0.4. In this context also, the pivotal angles are given by θ_{ij} for $|i-j| \leq \lambda$. The remaining angles are fixed to a constant value $\pi/2$.

3.2 Bayesian Variable Clustering

Clustering is a form of unsupervised learning where the objects are grouped on the basis of some similarity measures inherent among them. The interest and research on developing new clustering techniques have been proliferated owing to the emergence of several new disciplines which include but not limited to gene expression data in microarrays and portfolio analysis in finance. A vast and enriched literature of different clustering techniques have been developed in the last few decades in statistics, computer science and machine learning literature. The different algorithmic clustering techniques commonly used in practice are: (1) hierarchical clustering (agglomerative and divisive approach), (2) partition methods (K-means clustering) both of which hing on a distance metric (Bibby et al. (1979), Friedman et al. (2001), Rokach & Maimon (2005)) without assuming any underlying probability model for the clusters. (3) model-based approach which usually assumes a mixture model for the data. Recently, owing to the development of Bayesian non-parametric methods, different clustering algorithms based on Chinese Restaurant processes, Indian Buffet processes (Gershman & Blei, 2012), hierarchical Dirichlet processes (Teh et al., 2005; Kulis & Jordan, 2011) have been developed.

Explicating the pattern involved in a gene expression data or finance data is of utmost importance for proper understanding of the genomics factors in gene expression data and socio-ecomic factors influencing finance market respectively. However, the amount of data that one receives and underlying complexity of the pattern often pose challenges for interpretation and understanding

the results, necessitating a proper and meaningful clustering tool.

In this work, our aim is to cluster the variables which is drastically different from approaches for clustering observations or subjects . To understand it better, let $\mathbf{Y}$ denote a $n \times k$ data matrix consisting of k variables and n data points, represented in the matrix form

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & \cdots & y_{1k} \\ y_{21} & y_{22} & y_{23} & y_{24} & \cdots & y_{2k} \\ y_{31} & y_{32} & y_{33} & y_{34} & \cdots & y_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} & \cdots & y_{nk} \end{bmatrix} \tag{3.2}$$

From (3.2), one notes that each of n rows corresponds to one observation or data point whereas each of k columns pertains to one variable. A typical data clustering approach partitions the rows of $\mathbf{Y}$, i.e. essentially clustering of the observations. We are interested in partitioning the columns of $\mathbf{Y}$ which is essentially clustering of the variables, and correlations between the variables serve as our main building block to implement the algorithm. In a typical data clustering algorithm we consider how similar the objects are based on a similarity norm (say Euclidean or some other kind of distance). On the contrary, in a variable clustering problem, we are concerned with the correlation among the variables. Hence, highly correlated variables are more likely to lie in the same cluster.

Though there is a vast amount of works in the field of data clustering, but the variable clustering problem is at its infancy and has gotten limited attention (Bunea et al., 2018). In the absence of genuine variable clustering methods, very often traditional data clustering algorithms have been applied to this setup using brute force (Vigneau and Quannari, 2003; Duda et al., 2001), or ad hoc algorithms based on aspects of correlation matrices have been proposed. The literature on Bayesian methods for variable clustering is sparse with a few notable exceptions (Liechty et al. (2004), Palla et al. (2012)). Palla et al. (2012) developed a nonparametric Bayes algorithm based on Chinese restaurant process. On the other hand, our method is in the spirit of Liechty et al. (2004)

where a parametric model-based approach has been considered. A key advantage of our approach is that the number of clusters is unknown, and determined using a reversible jump Markov Chain Monte Carlo algorithm [RJMCMC](Green, 1995). The major obstacles in posterior sampling of the correlation parameters in the variable dimension Markov chain Monte Carlo algorithm are the maintenance of positive-definiteness constraint on the correlation matrix as well as computing the related normalizing constant.

In this thesis, our contributions are, (1) development of model-based Fraley & Raftery (2002) variable clustering method with different correlation structures, (2) proposing a novel variable clustering algorithm using the angular representation of the correlations which avoids some computational challenges due to the positive-definiteness constraint by using the Cholesky decomposition (Pinheiro & Bates, 1996; Rapisarda et al., 2007) and the ensuing angles (hyperspherical coordinates). We elicit substantive prior information on these angles which makes clustering of the variables feasible, (3) a data-driven estimate of number of clusters which traditional algorithms fail to provide. For the posterior inference, since the angle parameters are badly entangled in the posterior distribution, we resort to the Markov chain Monte Carlo algorithm (Tierney, 1994). For the posterior inference, we resort to the standard RJMCMC techniques as in (Green, 1995; Robert, 2004; Green & Hastie, 2009; Fan & Sisson, 2011).

3.3 Variable clustering: Introducing Prior models

3.3.1 Introduction

We consider n data $\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}$, where each $\mathbf{y_i}$ is a k-dimensional vector in R^k rendering to $\mathbf{y_i} = (y_{i1}, y_{i2}, \cdots, y_{ik})^\top$. In this set-up, y_{ij} corresponds to j-th variable in i-th data, where $j \in \{1, 2, \cdots, k\}$. Our goal is to cluster the variables which essentially implies seeking a partition $\mathcal{P} = \{1, 2, \cdots, m\}$ of $\{1, 2, \cdots, k\}$ based on the correlation values of the variables, assuming one variable belongs to one and only one cluster. In other words, we aim to partition the variables in

the following manner,

$$\mathbf{y_i} = (\underbrace{y_{i\sigma(1)}, y_{i\sigma(2)}, \cdots, y_{i\sigma(l_1)}}_{C_1}, \underbrace{y_{i\sigma(l_1+1)}, y_{i\sigma(l_1+2)}, \cdots, y_{i\sigma(l_2)}}_{C_2}, \cdots, \underbrace{y_{i\sigma(l_{m-1}+1)}, \cdots, y_{i\sigma(m)}}_{C_m})^\top, \tag{3.3}$$

where the correlation values of the variables belonging to a particular cluster are the same or nearly equal and higher than correlation with any variable belonging to a different cluster. In the above setting, $\sigma(.)$ denotes a permutation of $\{1, 2, ..., k\}$. Therefore, we start by standardizing the variables by their respective standard deviations so that covariance matrix of $\mathbf{y_i}$ will become a correlation matrix. With respect to (3.3) we define the clusters as $C_d = \{\{y_{ij}\} : corr(y_{ij}, y_{ij'}) = r_d\}$, where r_d is the cluster-specific correlation value, $d \in \{1, 2, \cdots, m\}$. Depending upon the partition $\mathcal{P}$, to each variable, say j-th variable, one can associate a latent vector $\mathbf{z_j} = (z_{j1}' z_{j2}, \cdots, z_{jm})^\top$, where z_{ju} takes the value 1 if the j-th variable belongs to C_u, the u-th cluster. Define a $k \times m$ matrix $\mathbf{Z}$ having rows $\mathbf{z_j}$, for $j = 1, 2, ..., k$.

From (3.3), two different correlation based models are possible up to permutations of rows and columns of the correlation matrix of $\mathbf{y_i}$.

(A) Block diagonal structure: This arises when we assume variables belonging to different clusters must have correlation zero, i.e. $Corr(y_{ij}, y_{ij'}) = 0$, if $y_{ij} \in C_u$ and $y_{ij'} \in C_v$ for $u \neq v$ and variables belonging to the same cluster have nearly equal correlation.

(B) Block common structure: This one assumes that correlation between any two variables belonging to the same cluster is the same. It is trivial to note that (A) arises as a special case of (B), when inter-cluster correlation is zero.

It is expected that in (A) and (B), variables which are grouped in the same cluster appear together. One can note that both of these models depend on the ordering of the variables. Permuting the variables will destroy these structures.

In a Bayesian model-based clustering algorithm, clusters are enforced through subtle role of prior model, following a likelihood model $L(\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}|R, \mathbf{Z}, m)$ of $\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}$ given correlation matrix R, partition induced indicator set $\mathbf{Z}$ and m. A cluster-inducing hierarchical prior

model comprises of (a) $p(m)$, (b) $p(\mathbf{Z}|m)$, and (c) $p(R|\mathbf{Z}, m)$. Within this hierarchical framework, we propose clustering prior models for (A) and (C) only in the correlation space. The same model in (C) can be used for (B) as well. Towards this end, we assume

$$m \sim truncPoisson(1; k) \tag{3.4}$$

$$\mathbf{z_j}|m \sim multinomial(k, (s_1, s_2, ..., s_m)^\top) \tag{3.5}$$

where $s_1 = s_2 = ... = s_m$ (assuming equal probabilities for each cluster), and k is the dimension which is known, where $truncPoisson(1; k)$ is a Poisson distribution upper truncated at k with parameter 1. With this set-up, clustering model in (A) and (C) can be calibrated with the following distribution for $p(R|\mathbf{Z}, m)$.

For block diagonal correlation (A), the joint prior on $R = (r_{ij})$ given $\mathbf{Z}, m$ could be of the form,

$$p(R|\mu, \sigma^2, \mathbf{Z}) = C(\mu, \sigma^2, \mathbf{z}) \prod_{i>j} \Big[\sum_{u=1}^{m} I(z_{iu} = 1) I(z_{ju} = 1) \text{exp}\{-(r_{ij} - \mu_u)^2 / 2\sigma_u^2\} \Big] I(R \in \mathcal{R}^k) \tag{3.6}$$

which translates to a $N(\mu_u, \sigma_u^2)$ prior for the correlation r_{ij} only if the i-th and j-th variable both belong to u-th cluster in the constrained space of k-dimensional correlation matrices, $\mathcal{R}^k$.

For a general correlation matrix in (C), one can use the following distribution of R given $\mathbf{Z}, m$ similar to the variable clustering prior from Liechty et al. (2004) of the form,

$$p(R|\mu, \sigma^2, \mathbf{Z}) = C(\mu, \sigma^2, \mathbf{Z}) \prod_{i>j} \Big[\sum_{u,u'} I(z_{iu} = 1) I(z_{ju'} = 1) \text{exp}\{-(r_{ij} - \mu_{uu'})^2 / 2\sigma_{uu'}^2\} \Big] I(R \in \mathcal{R}^k) \tag{3.7}$$

which essentially means a $N(\mu_{uu'}, \sigma_{dd'}^2)$ for correlation r_{ij} when i-th and j-th variable belong to C_u and $C_{u'}$ respectively. For the hyperparameters μ, one can assume a zero-mean normal distribution

with known variance and for σ^2, one can assume inverse-Gamma distribution with known scale and shape parameters.

The indicator function $I(R \in \mathcal{R}^k)$ ensures that R lies in the space of correlation matrices of order k ($\mathcal{R}^k$) enforcing the normalizing constant C to be a function of μ, σ^2 and $\mathbf{Z}$. Since block common correlation model (B) appears as a special case, the prior model in (3.7) is still applicable.

We do acknowledge that prior models in (3.6) and (3.7) are very intuitive and novel as far as modelling is concerned, but computationally very expensive for implementation. Due to the presence of indicator function, the posterior inference relies on Metropolis-Hastings (M-H) algorithm and a key step in that is to maintain positive definiteness and calculate normalizing constant C in every iteration.

With that preludial remark, we next revisit unconstrained angular reparameterization of a correlation matrix due to (Pinheiro & Bates, 1996; Rapisarda et al., 2007) for reader's convenience which often offers a flexible way of modelling covariance. We characterize block diagonal and block common correlation matrix with angles, enabling us to elicit modified priors on angles like (3.6) which differs substantially from a general structure.

3.3.2 Review of angular reparametrization (Θ) of R

This section describes connections between the hyperspherical coordinates (angles) and a correlation matrix $R = (r_{ij})$.

For a general $k \times k$ correlation matrix R with 1's in the diagonal, its Cholesky decomposition is given by $R = BB^\top$ where the Cholesky factor B is a lower triangular matrix. Since the rows of B are vectors of unit-length, it turns out that they admit the following representation involving

trigonometric functions of some angles (Pinheiro & Bates, 1996; Rapisarda et al., 2007):

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ c_{21} & s_{21} & 0 & 0 & \dots & 0 \\ c_{31} & c_{32}s_{31} & s_{32}s_{31} & 0 & \dots & 0 \\ c_{41} & c_{42}s_{41} & c_{43}s_{42}s_{41} & \prod_{j=1}^{3} s_{4j} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ c_{k1} & c_{k2}s_{k1} & c_{k3}s_{k2}s_{k1} & c_{k4}\prod_{j=1}^{3} s_{kj} & \dots & \prod_{j=1}^{k-1} s_{kj} \end{bmatrix} \tag{3.8}$$

with $c_{ij} = \cos(\theta_{ij})$ and $s_{ij} = \sin(\theta_{ij})$, where the angles θ_{ij}'s are measured in radians, $1 \leq j < i \leq k$. Restricting $\theta_{ij} \in [0, \pi)$ makes the diagonal entries of B non-negative, and hence B is unique to which we associate a $(k-1) \times (k-1)$ lower triangular matrix Θ with $k(k-1)/2$ angles:

$$\Theta = \begin{bmatrix} \theta_{21} & 0 & 0 & \dots & 0 \\ \theta_{31} & \theta_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \theta_{k2} & \theta_{k3} & \dots & \theta_{k,k-1} \end{bmatrix} \tag{3.9}$$

Note that the (i, j)-th element of Θ is denoted by $\theta_{i+1,j}$ so that θ_{ij} corresponds to the (i, j)-th element of R, we refer to Θ as the **angular matrix** associated to R. For further details and applications of these angles, see Creal et al. (2011), Zhang et al. (2015) and Tsay & Pourahmadi (2017). One can characterize (A) block diagonal and (B) Block common correlation matrices in terms of structured Θ matrix, which is completely determined by some (*pivotal*) angles.

3.3.3 Correspondence of clustering between R and Θ

Proposition 3.3.1. *For a block diagonal correlation matrix R = block diag$(R_1, R_2, \cdots, R_m)$, consisting of m equicorrelated blocks (r_i for block R_i), the corresponding angular matrix Θ is characterized by only m angles $\theta_1, \theta_2, ..., \theta_m$, where $r_i = cos\ \theta_i$.*

Proof. See appendix A.0.5. □

From Prop-block-diagonal, it follows that clustering among r_is is enforced by clustering of θ_is due to monotonicity of cosine function. Exploiting this fact, in our clustering algorithm later in this paper we cluster θ_is which will in turn cluster the variables based on correlations.
It follows immediately from Prop-block-diagonal that in case of block diagonal correlation matrix, clustering on correlations rendering to m different groups is equivalent to clustering of those m angles by the monotonicity of cosine function. However, this will be impose some conditions on the pivotal angles to maintain positive definiteness. Assuming each block has dimension k_i so that $\sum_{i=1}^{m} k_i = k$, the support of θ_i is $0 < \theta_i < \arccos(1/(k_i - 1))$ for $i = 1, 2, \cdots, m$.

Proposition 3.3.2. *Suppose that* $r_1 = \cos\theta_1$, $r_2 = \cos\theta_2$. *Then* $|\theta_1 - \theta_2| \geq \delta$ *if and only if* $|r_1 - r_2| \geq |1 - \cos\delta|$

Proof. See appendix A.0.6. □

3.3.4 Prior specification on the angles

Define a matrix $\mathbf{Z}_{k\times m}$ whose i-th row corresponds to the allocation of i-th variable in one of the m clusters, i.e.

$$\mathbf{Z}_{iu} = \begin{cases} 1 & \text{if } i\text{-th variable belongs to } u\text{-th cluster} \\ 0 & \text{otherwise} \end{cases}$$

Since we are assuming that a variable belongs to exactly one cluster, therefore, each row of $\mathbf{Z}$ contains exactly one 1 and rests are 0s. We assume the following hierarchical prior models for

$m, \mathbf{Z}, \Lambda$.

$$m \sim truncPois(m; 1, k) \tag{3.10}$$

$$\mathbf{Z}_i \sim Multinom(1; q_1, q_2, ..., q_m) \quad \text{for } i = 1, 2, ..., k \tag{3.11}$$

Having sampled $\mathbf{Z}$, the allocations are determined. Let k_u denote the size of u-th cluster, $u = 1, 2, ..., m$, i.e.

$$k_u = |\{i : z_{iu} = 1\}| \tag{3.12}$$

Then assume the following prior on $\theta_{piv} = (\theta_1, \theta_2, ..., \theta_m)^\top$ in order to shrink them to different values.

$$\theta_{piv}|\mathbf{Z}, m, \Lambda = \prod_{u=1}^{m} Q\Big(\theta_u; 0, \arccos\big(\frac{1}{k_u - 1}\big), \lambda_u\Big) \tag{3.13}$$

where $Q(\theta; 0, a, \lambda)$ is the density of truncated wrapped Exponential distribution between 0 and a with parameter λ. We are clustering the pivotal angles by introducing wrapped exponential distribution distribution with different parameters. We sample $\lambda_1, \lambda_2, \cdots, \lambda_m$ in the following manner,

$$\begin{aligned} &\lambda_1 \sim N^+(\lambda; 0, 1, 0, \infty) \\ &\lambda_2|\lambda_1 \sim N^+(\lambda; 0, 1, \lambda_1, \infty) \\ &\lambda_i|\lambda_{i-1} \sim N^+(\lambda; 0, 1, \lambda_{i-1}, \infty) \quad \text{for } i = 2, 3, \cdots, m \end{aligned} \tag{3.14}$$

,

where $N^+(; 0, 1, a, \infty)$ denotes a truncated positive normal distribution with $\mu = 0, \sigma = 1$ truncated between a and ∞. With the aforementioned prior specification, one notes the followings:

1. We have used truncated wrapped exponential distribution as the prior for the pivotal angles, since the mean of truncated wrapped exponential distribution has closed form expression. However, as an alternative, one can use any truncated circular distribution as prior for pivotal angles, for example von-Mises but the mean has no closed form expression.

2. From (3.13), it is noted that mean of θ_u is $\arctan(1/\lambda_u)$. It is evident from the joint prior on θ_{piv} that mean of the clusters are determined by λ_us. Therefore, to ensure that the non-overlapping support for pivotal angles which further renders to cluster separation, the hyper-priors on λ_us given in (3.14) is reasonable due to the to ordering among λ_us.

3.3.5 Cluster separability

Cluster separability is a fundamental challenge in any clustering algorithm. We are clustering the pivotal angles by introducing wrapped exponential distribution distribution with different parameters. It has been noted that the mean of wrapped exponential distribution is $\arctan(1/\lambda)$ where λ is the parameter of the wrapped exponential distribution. Since the mean of the clusters is related only to λs in the prior specification, we enforce cluster separability by ordering the values of λs in the prior model through following specification. where $N^+(;0,1,a,\infty)$ denotes a truncated positive normal distribution with $\mu = 0, \sigma = 1$ truncated between a and ∞. Note that λ_is generated above satisfy $\lambda_1 < \lambda_2 < \lambda_2 < ... < \lambda_m$.

3.4 Posterior computation

The posterior distribution is given by

$$p(\Theta, \mathbf{Z}, \Lambda, m|\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}) \propto L(\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}|\Theta, \mathbf{Z}, m) \times p(m) \times p(\Theta|\Lambda, \mathbf{z}, m) \times p(\Lambda) \times p(\mathbf{Z}|m) \tag{3.15}$$

Our goal in this section is to first estimate number of clusters m and posterior of R and Λ. Once this has been determined we can obtain the allocation of the variables in the corresponding block by using the full conditional distribution of z_{iu}. The whole algorithm is, thus, accomplished in two

steps performing a reversible jump Markov chain Monte Carlo(RJMCMC) algorithm since we are assuming that the number of blocks m is not known apriori followed by a Monte Carlo approach to calculate allocation probabilities of the variables in the clusters. We describe posterior sampling scheme in detail for general correlation prior $(\mathbf{I})$. For block diagonal model, computation follows similarly.

3.4.1 Step1: Sampling Λ, R

From proposed priors, one can note that the clusters are induced by the elements of Λ, thus, in the following RJMCMC algorithm (Green, 1995; Robert, 2004; Green & Hastie, 2009; Fan & Sisson, 2011), one element of Λ, say λ_j is randomly split into $(\lambda_{j_1}, \lambda_{j_2})$ and then two elements of Λ are merged into a single element. The algorithm is summarized as follows:

- 1. Initialize Θ, Λ. In the initialization step, one may assume one block common structure (AR(1) matrix) to initialize Θ.
- 2. Birth step:

  Split λ_j to $(\lambda_{j_1}, \lambda_{j_2})^\top$ by $\lambda_{j_1} = \lambda_j + \tau$, $\lambda_{j_2} = \lambda_j - \tau$, where $\tau \sim Unif(-\pi/4, \pi/4)$.

  Acceptance probability $\alpha = \min\{1, \frac{p(\Theta, \lambda_{j_1}, \lambda_{j_2}, d(j_1, j_2))}{p(\Theta, \lambda_j, d(j))} \times \frac{2}{\pi} \times |\frac{\partial(\lambda_{j_1}, \lambda_{j_2})}{\partial(\lambda_j, \tau)}|\}$
- 3. Death step: Two components λ_{j_1} and λ_{j_2} are merged to a single component $\lambda_j = (\lambda_{j_1} - \tau + \lambda_{j_2} + \tau)/2$ with acceptance probability $\alpha = \min\{1, \frac{p(\Theta, \lambda_j, d(j))}{p(\Theta, \lambda_{j_1}, \lambda_{j_2}, d(j_1, j_2))} \times \frac{\pi}{2} \times |\frac{\partial(\lambda_j, \tau)}{\partial(\lambda_{j_1}, \lambda_{j_2})}|\}$
- 4. Step 1, 2 and 3 are repeated as many times as required and the value of m is determined by which stage is visited maximum number of times and posterior estimate of R is obtained by averaging over those stages.

3.4.2 Step 2: Sampling from the full conditional distribution of z_i

The full conditional distribution of z_i is

$$p(z_{iu} = 1|.) \propto \prod_{j \neq i} q(\theta_{ij}; \lambda_{u,v_{z_j}}) \tag{3.16}$$

Therefore, posterior estimate of z_{iu} which corresponds to the allocation probabilities that i-th variable is in u-th cluster, is obtained by

$$\hat{p}(z_{iu} = 1|.) = \frac{\prod_{j \neq i} q(\hat{\theta}_{ij}; \hat{\lambda}_{u,v_{z_j}})}{\sum_{u' \neq u} \prod_{j \neq i} q(\hat{\theta}_{ij}; \hat{\lambda}_{u',v_{z_j}})}$$

3.5 Simulations and Data Analyses

In this section, we compare numerical performance of our Bayesian Variable Clustering (BVC) algorithm with a recent method based on COD (COvariance Difference) of Bunea et al. (2018) and the classical or standard K-means clustering algorithm. The performance criterion we use is the proportion of true recovery which is defined for a k-dimensional correlation matrix consisting of k variables as

$$\frac{\#\text{variables in the true clusters}}{k}.$$

COD has been implemented using the R package “cord" available in CRAN.R-project.org and K-means algorithm has been implemented on the transposed data matrix avialble in “kmeans" function in R software.

3.5.1 Simulation design M1S of Bunea et al. (2018)

In this section, we compare numerical performance of our Bayesian Variable Clustering (BVC) algorithm with a recent method based on COD (Covariance Difference) of Bunea et al. (2018), Partitioning Around Medoids (PAM) algorithm which minimizes the Manhattan distance of the data points to the medoids (Kaufman & Rousseeuw, 2009) and the classical or standard K-means clustering algorithm. The performance criterion we use is the proportion of true recovery which is

defined for a k-dimensional correlation matrix consisting of k variables as

$$\frac{\#\text{ of variables in the true clusters}}{k}.$$

COD and PAM have been implemented using the R packages cord and class available via CRAN and K-means algorithm has been implemented on the transposed data matrix available in *kmeans* function in R software.

3.5.2 Simulation study

In this simulation experiment the setup is that of the model M1 in Bunea et al. (2018), where we start with an $m \times m$ matrix $C = B^\top B$ where the entries of the random $(m-1) \times m$ matrix B take values $-1, 0, 1$ with probabilities $0.5 \times m^{-1/2}$, $1 - m^{-1/2}$ and $0.5 \times m^{-1/2}$, respectively, with m being the number of clusters. Next, we consider a balanced case with each group (cluster) of size k/m. Let $A = (a_{ij})$ be the $k \times m$ membership matrix with $a_{ij} = 1$ if the i-th variable belongs to C_j and 0 otherwise. Finally, consider the covariance matrix $\Sigma = ACA^\top + \Gamma$ where Γ is a diagonal matrix whose entries are random permutations of $\{0.5, 0.5 + 1.5/(k-1), ..., 2\}$ and the corresponding correlation matrix R. With $k = 200, m = 4$, we simulate n independent observations from a multivariate normal distribution with mean zero vector and covariance matrix R. We consider four different sample sizes $n = 100, 300, 600, 900$ to compare BVC, COD and K-means algorithms with respect to cluster recovery. The results presented in Figure 3.1 shows the superior performance of BVC relative to COD, PAM and K-means.

3.5.3 Application of Protein clustering to Hereditary Breast Cancer Data

Breast cancer is one of the most common cancers with a massive number of cases reported. For instance, in 2018, more than 268,000 Americans were estimated to have been diagnosed and 41,000 were estimated to have died from breast cancer related tumors (Siegel et al., 2018). The Cancer Genome Atlas: TCGA is the largest available cancer data consortium consisting of parallel mRNA expressions, DNA copy number, methylation expressions, protein expressions, along with clinical variables such as survival or the tumor stages for a total of 33 types of tumors. Among

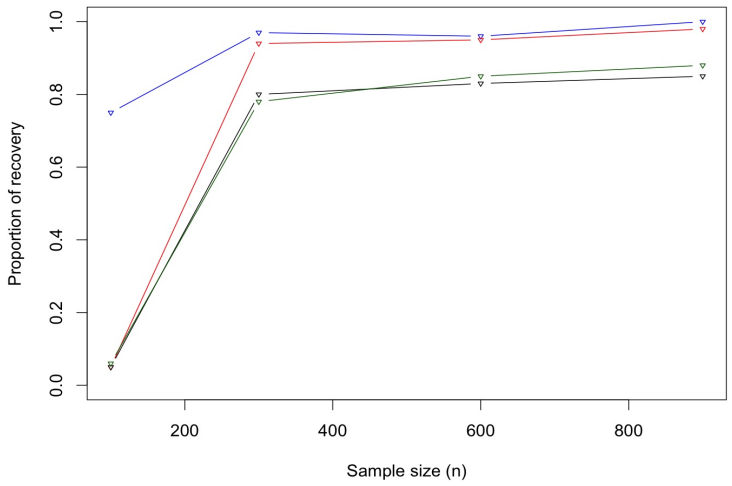

Figure 3.1: Comparing BVC(blue), COD(red), PAM(green) and K-means(black) for simulation study in M1S

them we consider the information of 222 breast tumor samples; we consider 27 different proteins 4 different pathways (see 3.7.1).

Applying our BVC algorithm to this data, the MAP estimate of the number of clusters is 4, which is consistent with the number of pathways. However, applying the COD algorithm in Bunea et al. (2018) the estimated number of clusters is 23, much larger than the known value of 4. In Table 3.2, we provide the assignments of various proteins in different clusters. Additionally, for the sake of comparison we have also applied the K-means algorithm to this data for $k = 4, 23$, respectively, with results reported in Table 3.2. The results suggest that our Bayesian variable clustering (BVC) is performing better to cluster the proteins with respect to pathways. Only misclassified proteins are $MAPK\_pT201\_Y204$, $CD31$, $CD49b$, $CDK1$. The COD algorithm reports that number of clusters is 23 which appears to be too high since the number of proteins is 27. A possible reason could be this algorithm is meant for high dimensional clustering, it fails to detect clustering configuration in small dimensional cases. Comparisons with standard K-means and PAM algorithm also reveal that these two methods result in more disagreement of the cluster configuration

of the proteins according to the pathway information. This apart, K-means and PAM algorithm disagree among themselves, e.g., $ER.alpha$, JNK_pT183_pT185 etc.(Table 3.2). We have also performed hierarchical clustering on this data with various linkages . The results are presented in Figure 3.2.

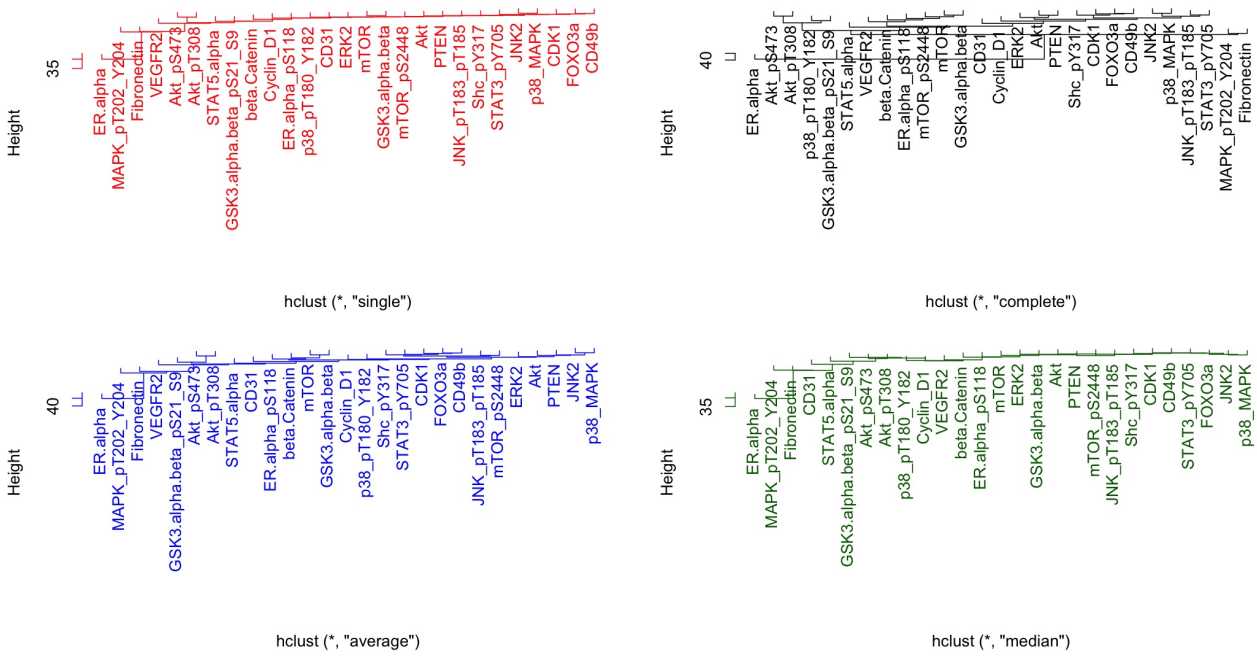

Figure 3.2: Hierarchical clustering for the protein expression data with four different linkages.

To assess the performance of the proposed method with increasing sample size and under different correlation structure, we perform further simulation study ($S1$-$S4$). The starting point is the posterior estimate of the correlation matrix R_{post} from BVC which is a block diagonal matrix with entries 0.33, 0.77, 0.15 and 0.25. We simulate a sample of size n from a multivariate normal distribution having mean zero vector and covariance equals to R_{post} or some indicated modifications of it provided the matrix is a valid correlation matrix, with n varying over $\{50, 100, 300, 500, 1000\}$. In the following we provide the details of the experiments.

S1. In S1, R_{post} has been used as population covariance to generate samples. In Figure 3.3 we

depict the proportion of times the true clusters are recovered by the different methods and this is done for different n. Figure 3.3 (S1) suggests the better performance of BVC and COD compared to K-means and PAM. It is to be noted here that for small sample size ($n = 50$), performance of BVC is better than COD. However, with the increasing sample size, performance of BVC and COD are approximately same.

S2. Here, we consider an approximate block-diagonal correlation matrix R of order 27 with four different clusters $\{C_1, C_2, C_3, C_4\}$, where correlations in the blocks are 0.33 (C_1), 0.77 (C_2), 0.05 (C_3), 0.25 (C_4)(same as R_{post}), and additionally those in the off-diagonal blocks are 0.44 (C_1, C_2), 0.19 (C_1, C_3), 0.29 (C_1, C_4), 0.4 (C_2, C_3), 0.42 (C_2, C_4), 0.18 (C_3, C_4). It follows from Figure 3.3 (S2) that BVC outperforms all its competitors for $n = 50, 100, 300, 500$. For $n = 1000$, performance of BVC and COD are the same.

S3. In this setup, we allow the variables in C_3 and C_4 to be minimally separated by changing the correlation in C_3 from 0.05 (under S2) to 0.22 which is closer to the fourth cluster. The results plotted in Figure 3.3 (S3) confirm the degradation of the performance of all four methods.

S4. The final modification corresponds to changing the off-diagonal block entries where we replace the values 0.44 (C_1, C_2), 0.19 (C_1, C_3), 0.29 (C_1, C_4) by the same value 0.44 in the first row- and column-blocks and correlation between (C_3, C_4) is set to 0.36. The results are reported in Figure 3.3 (S4). All the algorithms suggest that there is exactly one cluster irrespective of the sample size.

3.5.4 Finance Data

We consider daily log returns of five financial and two industrial stocks from January 3, 2000 to December 31, 2009 for 2515 observations downloaded from Center for Research in Security Prices (CRSP)(Tsay & Pourahmadi, 2017). It is known that the stocks belong to three industry groups; Group 1 contains Morgan Stanley (MS) and Goldman Sachs (GS) which are investment

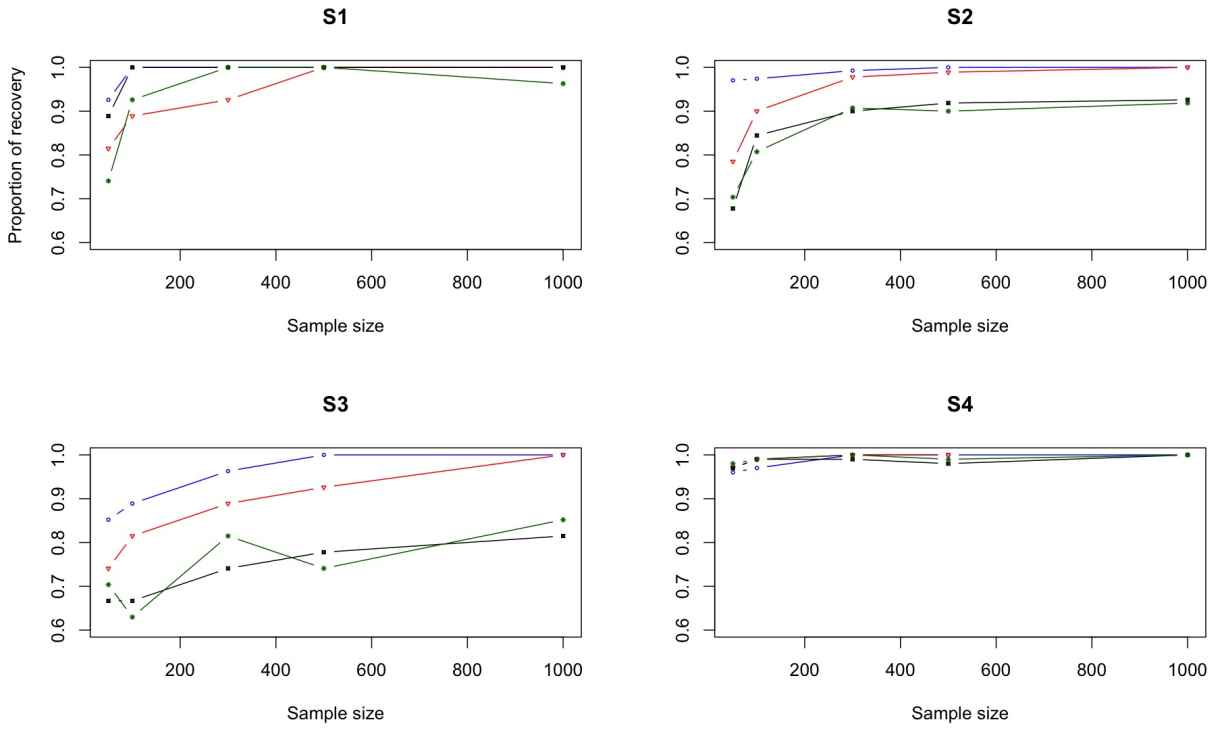

Figure 3.3: The proportion of times the true clusters are recovered by BVC (blue), COD (red), K-means (black) and PAM (green) against different sample sizes for studies S1-S4.

bankers; Group 2 contains Bank of America (BAC), J.P. Morgan Chase (JPM), and Wells Fargo (WFC) which are retail bankers and Group 3 contains Boeing (BA) and Intel (INTC) Corporation.

Implementing our BVC on this data with $n = 2515$ and $k = 7$, it suggests that number of clusters is 3 and 2 with posterior probabilities 0.79 and 0.21 respectively. This supports the fact that it is reasonable to cluster the companies based on their business sectors. We provide the posterior probabilities of lying in either of the two clusters of the companies in Table 3.1 and proportion of recovery in Figure 3.4.

3.6 Discussion

We have proposed a correlation matrix based Bayesian clustering technique to recover the protein signaling pathways. This method uses angular reparameterization of correlation matrix with the specification of wrapped exponential prior on the angle parameters. Nonetheless, as an

Table 3.1: Clustering posterior probabilities of companies

Company	C_1	C_2	C_3
MS	**0.61**	0.30	0.09
GS	**0.55**	0.40	0.05
BAC	0.20	**0.65**	0.15
JPM	0.15	**0.55**	0.40
WFC	0.15	**0.60**	0.25
BA	0.15	0.20	**0.65**
INTC	0.15	0.15	**0.70**

alternative, one can use any truncated circular distribution as prior for pivotal angles, for example von-Mises distribution. However, this particular choice produces a mean which has no closed form and as a result our proposed method can not be carried out for a posterior analysis.

A large amount of recent interest is being channelized to analyze the proteomics data directly because direct analysis of proteins has potential to uncover the cell functional characteristics. When it is of interest to find the group of proteins having similar functions which may be evident via their expression measurements then our proposed method can be used to bridge that gap. As mentioned earlier, our method is particularly useful when the number of clusters is not known and hence is learned via the posterior MCMC, which is often the case for the real data where the determining the number of clusters is itself a tedious job.

3.7 Pathways and Cluster Assignments of Proteins

3.7.1 Pathway Information

(1) MAP kinase pathway has 8 proteins: $ER.alpha$, $ER.alpha_pS118$, $ERK2$, $JNK2$, JNK_pT183_pT185, $MAPK_pT202_Y204$, $p38_MAPK$, $p38_pT180_Y182$;

(2) $PI3K/AKT/mTOR$ signaling pathway has 7 different proteins: Akt, Akt_pS473, Akt_pT308, $FOXO3a$, $PTEN$, $mTOR$, $mTOR_pS2448$;

(3) JAK-STAT Signaling pathway has 3 different proteins: Shc_pY317, $STAT3_pY705$, $STAT5.alpha$;

(4) Wnt signaling pathway has 9 different proteins: $CD31$, $CD49b$, $CDK1$, $Cyclin_D1$, $Fibronectin$,

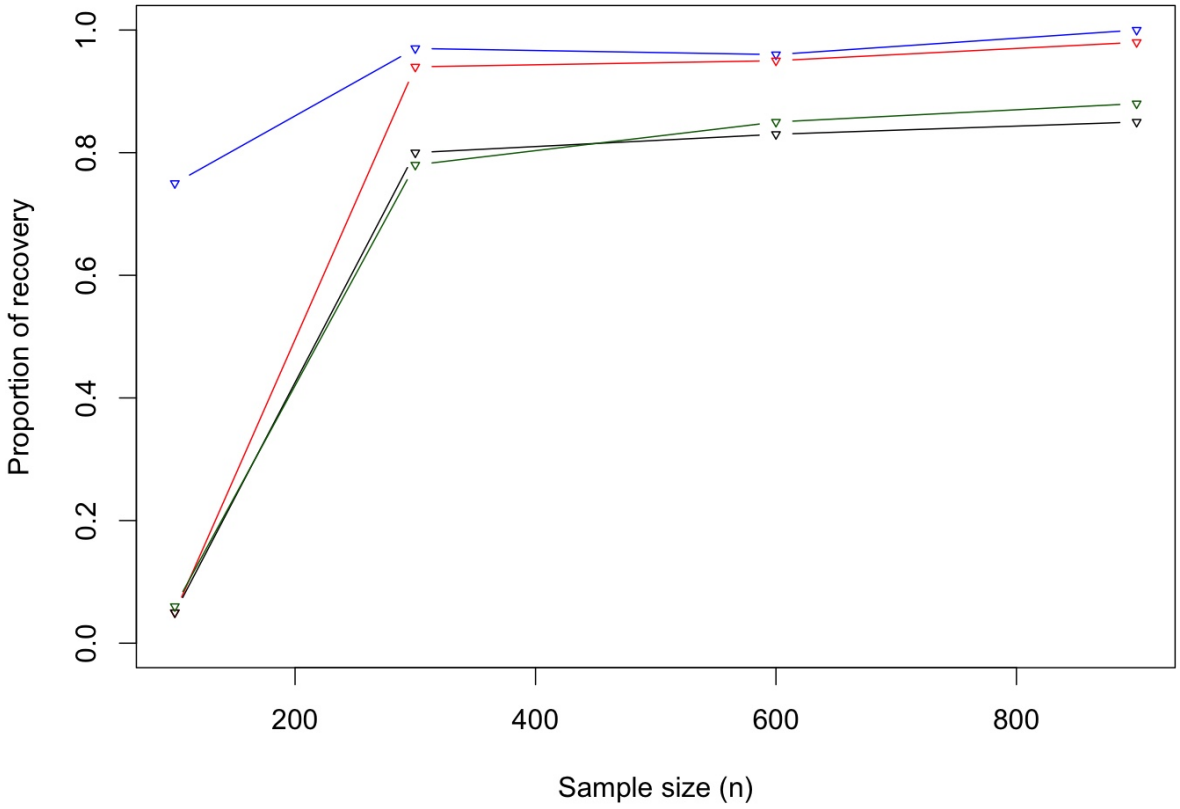

Figure 3.4: Comparing BVC(blue), COD(red) and K-means(black) for M1S(left) and finance data with 100 iterations.

$GSK3.alpha.beta$, $GSK3.alpha.beta_pS21_S9$, $VEGFR2$, $beta.Catenin$.

3.7.2 Cluster Assignments of Proteins

Table 3.2 presents the cluster assignments of proteins by BVC, COD and K-means algorithms.

Table 3.2: Cluster comparisons by BVC, COD and K-means

Protein	BVC	COD	K-means (k=4)	K-means (k=23)
$ER.alpha$	C_1	C_1	C_3	C_{17}
$ER.alpha_pS118$	C_1	C_1	C_4	C_{16}
$ERK2$	C_1	C_2	C_4	C_2
$JNK2$	C_1	C_5	C_1	C_8
JNK_pT183_pT185	C_1	C_6	C_1	C_4
$MAPK_pT202_Y204$	C_3	C_7	C_2	C_{12}
$MAPK_pT202_Y204$	C_3	C_8	C_1	C_8
$p38_MAPK$	C_1	C_9	C_1	C_{20}
$p38_pT180_Y182$	C_1	C_{10}	C_1	C_{11}
Akt	C_2	C_{11}	C_2	C_{19}
Akt_pS473	C_2	C_{12}	C_2	C_{10}
Akt_pT308	C_2	C_{13}	C_1	C_3
$FOXO3a$	C_2	C_{14}	C_1	C_6
$PTEN$	C_2	C_{15}	C_4	C_{13}
$mTOR$	C_2	C_{16}	C_1	C_5
$mTOR_pS2448$	C_2	C_{17}	C_1	C_{15}
Shc_pY317	C_3	C_{18}	C_1	C_{18}
$STAT3_pY705$	C_3	C_2	C_4	C_{22}
$STAT5.alpha$	C_3	C_3	C_1	C_7
$CD31$	C_2	C_{19}	C_1	C_3
$CD49b$	C_2	C_3	C_1	C_3
$CDK1$	C_2	C_{20}	C_1	C_{21}
$Cyclin_D1$	C_4	C_{21}	C_1	C_9
$Fibronectin$	C_4	C_4	C_4	C_{13}
$GSK3.alpha.beta$	C_4	C_{22}	C_2	C_{14}
$GSK3.alpha.beta$	C_4	C_{23}	C_4	C_1
$GSK3.alpha.beta_pS21_S9$	C_4	C_4	C_4	C_{23}

4. CONCLUSIONS AND FUTURE RESEARCH

The major aspect of this dissertation is to study angular parametrization of a correlation matrix by using its Cholesky decomposition and explore its certain characteristics. In particular, we have provided the statistical interpretation or meaning of the angles (arising out of this parameterization) for the first time which are inverse cosine of the semi-partial correlation (SPC)s. In Chapter 2, this parameterization is exploited in Bayesian estimation of correlation matrices in the context of longitudinal data. We have shown that this expedites posterior computation by avoiding the positive definiteness constraint in an iterated model fitting procedure. Comparisons have been made with constrained approach (Barnard et al. (2000)) and a recent partial autocorrelation based approach Gaskins et al. (2014) which our method outperforms others in terms of time complexity and posterior risk. In Chapter 3, we have characterized some structured correlation matrices of special interest by angles and provided a method to perform Bayesian variable clustering based on block diagonal with equicorrelated block structure. Starting with unknown number of clusters, our reversible jump Markov chain Monte Carlo based algorithm estimates number of clusters and provides clustering configurations of the variables in the posterior computation. We have shown superior performance of our approach by comparing with traditional clustering including K-means and one method based on covariance difference by Bunea et al. (2018). In summary, we have studied two aspects where this angle based parameterization can be suitably adapted. However, there are some potential aspects where the application of this approach might be investigated. We end this dissertation by stating two of them.

1. Probabilistic aggregation is a fundamental problem in Geology. The USGS National Assessment of Geologic Carbon Dioxide Storage generates the probability distribution of CO_2 that can be stored in subsurface rock units in a supercritical state by Monte Carlo simulation of a probabilistic model of storage. The assessment focuses on existing pore space in saline formations beneath a regional seal, including but not limited to depleted hydrocarbon reservoirs. In each basin, porous

reservoirs with differing characteristics are identified and designated as storage assessment units (SAUs). The challenge is to get the aggregated distribution of CO_2 on the national level from user-specified dependencies. To present the problem statistically, consider m basins designated by $X_1, X_2, \cdots, X_m$ where the l^{th} basin consists of a total of n_l SAUs, $l = 1, 2, \cdots, m$. Let X_{il} be the carbon dioxide storage (in megatons) in the i^{th} SAU of l^{th} basin, $i = 1, 2, \cdots, n_l$, $l = 1, 2, \cdots, m$.

With notation as above, each X_l is a vector of dimension $n_l \times 1$. For the time being, we assume a multivariate normal distribution for the partitioned vector $X = (X_1, X_2, \cdots, X_m)$, with mean vector $(\mu_1, \mu_2, \cdots, \mu_m)$, $\mu_l = E(X_l)$ and covariance matrix

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_m) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_m) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(X_m, X_1) & Cov(X_m, X_2) & \dots & Var(X_m) \end{bmatrix} \tag{4.1}$$

Then, one is typically interested in finding the *aggregate distribution* (over the basins) of

$$S = \sum_{l=1}^{m} \sum_{i=1}^{n_l} X_{il},$$

and its variance

$$Var(S) = \sum_{i=1}^{m} \sum_{l=1}^{n_l} Var(X_{il}) + \sum_{i=1}^{m} \sum_{l \neq r} Cov(X_{il}, X_{ir}) + \sum_{i \neq j} \sum_{l,r} Cov(X_{il}, X_{jr}). \tag{4.2}$$

Note that in the variance formula, the second and third summands contribute to within and between basin covariances, respectively.

Finding the aggregate distribution requires knowing the marginal distribution of each X_{ij} and correlations. These are usually not available, but can be elicited by expert judgement specifying dependencies between pairs of SAUs, then the challenge lies in ensuring that Σ to be a valid covariance matrix or the corresponding R to be a valid correlation matrix. The dependency among SAUs arises because of several geological factors, e.g. same rock can source adjacent assessment

units; human factors due to assessors.

The aggregation algorithm that is commonly in practice Blondes et al. (2013) uses an approximate algorithm of computing a nearest correlation matrix of Higham (2002) which could destroy a particular desired structure. The key steps of this algorithm are:

- Initially a proto-correlation matrix $\tilde{R}$ of order k is specified by experts.
- If $\tilde{R}$ is not a valid correlation matrix, then it is replaced by the nearest correlation matrix $R = ((r_{ij}))$ as in Higham (2002).
- A $n \times k$ matrix M of sample numbers is generated following an Algorithm (Schuenemeyer & Gautier (2010),Blondes et al. (2013)) with a correlation structure within the sampling error of R, where n is the number of trials of k-dimensional data points $Y = (y_1, y_2, \cdots, y_k)^\top$.

One can note that correcting $\tilde{R}$ to a nearest correlation matrix can destroy a particular structure inherent in the study. Since this angle based approach guarantees a valid correlation matrix, the corresponding structured angular matrix should be investigated to resolve this problem.

2. Simulating sample correlation matrices is an important issue in many areas of statistics. Approaches such as generating Gaussian data and finding their sample correlation matrix or generating random uniform $[-1, 1]$ deviates as pairwise correlations both have drawbacks Hardin et al. (2013) in terms of computing the sample correlation matrix and find the difference between the estimate and the template; histograms of those differences and the distribution of the correlation error terms. In this context, Hardin et al. (2013) provided algorithms tailored to constant correlation structure, Toeplitz correlation structure, Hub correlation structure and a general structure to add noise to a correlation structure and showed that their algorithm performs better than generating Gaussian data in terms of the aforementioned criteria. However, the computational complexity of

their algorithm needs to be investigated. Since the angular reparametrization is capable of generating random correlation matrices Pourahmadi & Wang (2015), the corresponding angle based approach is still an open problem.

REFERENCES

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley-Interscience.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 1281–1311.

Bernardo, J. M., & Smith, A. F. (2001). *Bayesian theory.* IOP Publishing.

Bibby, J., Kent, J., & Mardia, K. (1979). *Multivariate Analysis.* Academic Press, London.

Blondes, M. S., Schuenemeyer, J. H., Olea, R. A., & Drew, L. J. (2013). Aggregation of carbon dioxide sequestration storage assessment units. *Stochastic Environmental Research and Risk Assessment*, *27*(8), 1839–1859.

Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). John Wiley & Sons.

Brown, P. (2002). The generalized inverted wishart distribution. *Encyclopaedia of Environmetrics. Wiley, England*, 1079–1083.

Brown, P. J., Le, N. D., & Zidek, J. V. (1994). Inference for a covariance matrix. *Aspects of uncertainty: a tribute to DV Lindley*, 77–92.

Bunea, F., Giraud, C., Luo, X., Royer, M., & Verzelen, N. (2018). Model Assisted Variable Clustering: Minimax-optimal Recovery and Algorithms. arxiv preprint arxiv:1508.01939v4.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

Computation, M., Golub, G., & Van Loan, C. (1996). Johns hopkins univ. *Press, London*.

Cooke, R. M., Joe, H., & Aas, K. (2011). Vines arise. *Dependence Modeling: Vine Copula Handbook*, 37–71.

Creal, D., Koopman, S. J., & Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, *29*(4), 552–563.

Daniels, M., & Pourahmadi, M. (2002). Dynamic models and bayesian analysis of covariance matrices in longitudinal data. *Biometrika*, *89*, 553–566.

Daniels, M. J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *Journal of Statistical Planning and Inference*, *127*(1-2), 119–130.

Daniels, M. J., & Kass, R. E. (1999). Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, *94*(448), 1254–1263.

Daniels, M. J., & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, *57*(4), 1173–1184.

Daniels, M. J., & Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, *89*(3), 553–566.

Dey, D. K., Srinivasan, C., et al. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, *13*(4), 1581–1591.

Eaves, D., & Chang, T. (1992). Priors for ordered conditional variance and vector partial correlation. *Journal of Multivariate Analysis*, *41*(1), 43–55.

Fan, Y., & Sisson, S. A. (2011). Reversible jump MCMC. *Handbook of Markov Chain Monte Carlo*, 67–92.

Fox, E., & Dunson, D. (2011). Bayesian nonparametric covariance regression. *arXiv preprint arXiv:1101.2017*.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611–631.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1). Springer Series in Statistics New York.

Gaskins, J., Daniels, M., & Marcus, B. (2014). Sparsity inducing prior distributions for correlation matrices of longitudinal data. *Journal of Computational and Graphical Statistics*, *23*(4), 966–984.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian*

data analysis (Vol. 2). CRC press Boca Raton, FL.

Gelman, A., et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Green, P. J., & Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, *155*(3), 1391–1403.

Haff, L., et al. (1991). The variational form of certain bayes estimators. *The Annals of Statistics*, *19*(3), 1163–1190.

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 1733–1762.

Higham, N. J. (2002). Computing the nearest correlation matrix?a problem from finance. *IMA journal of Numerical Analysis*, *22*(3), 329–343.

Hoff, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(5), 971–992.

Huang, A., Wand, M. P., et al. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439–452.

Huber, J. (1981). Partial and semipartial correlation-a vector approach. *The Two-Year College Mathematics Journal*, *12*(2), 151–153.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, 730–773.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10), 2177–2189.

Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, *104*(486), 682–693.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Kulis, B., & Jordan, M. I. (2011). Revisiting k-means: New algorithms via Bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*.

Lan, S., Holbrook, A., Fortin, N. J., Hernando, O., & Shahbaba, B. (2017). Flexible bayesian dynamic modeling of covariance and correlation matrices. *arXiv preprint arXiv:1711.02869*.

Leonard, T., Hsu, J. S., et al. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, *20*(4), 1669–1696.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, *100*(9), 1989–2001.

Liechty, J. C., Liechty, M. W., & Müller, P. (2004). Bayesian correlation estimation. *Biometrika*, *91*(1), 1–14.

Lin, S. P. (1985). A monte carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, 411–429.

Liu, X., Daniels, M. J., & Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association*, *104*(486), 429–438.

Madar, V. (2015). Direct formulation to cholesky decomposition of a general nonsingular correlation matrix. *Statistics & probability letters*, *103*, 142–147.

Marcus, B. H., Albrecht, A. E., King, T. K., Parisi, A. F., Pinto, B. M., Roberts, M., ... Abrams, D. B. (1999). The efficacy of exercise as an aid for smoking cessation in women: a randomized controlled trial. *Archives of internal medicine*, *159*(11), 1229–1234.

Mardia, K. V., & Jupp, P. E. (2009). *Directional statistics* (Vol. 494). John Wiley & Sons.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Palla, K., Ghahramani, Z., & Knowles, D. A. (2012). A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems* (pp. 2987–2995).

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, *6*(3), 289–296.

Pourahmadi, M., & Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the cholesky factor. *Statistics & Probability Letters*, *106*, 5–12.

Press, S. J. (1982). Applied multivariate analysis: Using bayesian and frequenist methods of inference. *Krieger Publishing Co., Florida*, *12*, 58–79.

Rapisarda, F., Brigo, D., & Mercurio, F. (2007). Parameterizing correlations: a geometric interpretation. *IMA Journal of Management Mathematics*, *18*(1), 55–73.

Robert, C. P. (2004). *Monte Carlo methods*. Wiley Online Library.

Roberts, G. O., & Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, *18*(2), 349–367.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321–352). Springer.

Schuenemeyer, J. H., & Gautier, D. L. (2010). Aggregation methodology for the circum-arctic resource appraisal. *Mathematical Geosciences*, *42*(5), 583–594.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, *68*(1), 7 – 30.

Smith, M., & Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, *97*(460), 1141–1153.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stein, C. (1956). *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution* (Tech. Rep.). STANFORD UNIVERSITY STANFORD United States.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing Clusters among Related Groups: Hierarchical dirichlet Processes. In *Advances in Neural Information Processing Systems* (pp. 1385–1392).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.

Tsay, R. S., & Pourahmadi, M. (2017). Modelling structured correlation matrices. *Biometrika*, *104*(1), 237–242.

Wang, H., & Pillai, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *Journal of Computational and Graphical Statistics*, *22*(3), 689–707.

West, M., & Escobar, M. D. (1993). *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University.

Wilson, A. G., & Ghahramani, Z. (2010). Generalised wishart processes. *arXiv preprint arXiv:1101.0240*.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.

Wong, F., Carter, C. K., & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, *90*(4), 809–830.

Yang, R., & Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 1195–1211.

Zhang, W., Leng, C., & Tang, C. Y. (2015). A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(1), 219–238.

APPENDIX

FIRST APPENDIX

A.1 Proof of Theorem 2.2.1

Before proving Theorem 1, we state following two useful propositions required for the proof.

Proposition A.1.1. *The correlation coefficient r_{ij} is related to the semi-partial correlations via*

$$r_{ij} = \rho_{ji:1,2,\ldots,j-1} \prod_{u=1}^{j-1} \sqrt{1-\rho^2_{uj:1,2,\ldots,u-1}} \sqrt{1-\rho^2_{ui:1,2,\ldots,u-1}} + \sum_{u=1}^{j-1} \left(\rho_{uj:1,2,\ldots,u-1} \prod_{m=1}^{u-1} \sqrt{1-\rho^2_{mj:1,2,\ldots,m-1}} + \rho_{ui:1,2,\ldots,u-1} \prod_{m=1}^{u-1} \sqrt{1-\rho^2_{mi:1,2,\ldots,m-1}} \right) \textit{for } 1 \leq j < i \leq k. \tag{A.1}$$

Proof. The proof is based on the following recurrence relation for partial correlations (Lewandowski et al., 2009),

$$\rho_{ji:k,L} = \frac{\rho_{ji:L} - \rho_{ki:L}\rho_{kj:L}}{\sqrt{(1-\rho^2_{ki:L})(1-\rho^2_{kj:L})}} \tag{A.2}$$

where L is a possibly empty set of indices with $i, j, k \notin L$. Using the above recurrence relation repeatedly, one notes that

$$r_{ji} = \rho_{ji} = \rho_{ji:1}\sqrt{(1-\rho^2_{1j})(1-\rho^2_{1i})} + \rho_{1j}\rho_{1i}$$

$$\rho_{ji:1} = \rho_{ji:1,2}\sqrt{(1-\rho^2_{2j:1})(1-\rho^2_{2i:1})} + \rho_{2j:1}\rho_{2i:1}$$

Therefore, $r_{ji} = \rho_{ji:1,2}\sqrt{(1-\rho^2_{2j:1})(1-\rho^2_{2i:1})(1-\rho^2_{1j})(1-\rho^2_{1i})} + \rho_{2j:1}\rho_{2i:1}\rho_{1j}\rho_{1i}$

which is conformable with (27). One can further substitute $\rho_{ji:1,2}$ by terms involving $\rho_{ji:1,2,3}$ and using these repeatedly the result follows. □

Proposition A.1.2. *If* $b_{m,j} = \rho_{jm:1,2,...,j-1} \prod_{u=1}^{j-1} \sqrt{1 - \rho^2_{um:1,2,...,u-1}}$, *then*

$$1 - \sum_{u=1}^{j-1} b^2_{ju} = \prod_{u=1}^{j-1} (1 - \rho^2_{uj:1,2,...,u-1}). \tag{A.3}$$

Proof. We first note that number of terms in left-hand side in (A.3) and right-hand side in is j. Then the proof will be complete by comparing the coefficient of $\rho^2_{mj:1,2,...m-1}$ on both sides in (A.3) where m can take any value in $\{1, 2, ..., j-1\}$.

Note that $\rho^2_{mj:1,2,...,m-1}$ enters in left-hand side of (A.3) only through b^2_{jm} and thus its coefficient equals to $-\prod_{u=1}^{m-1} \sqrt{1 - \rho^2_{um:1,2,...,u-1}}$ which is also the coefficient of $\rho^2_{mj:1,2,...,m-1}$ in right-hand side. □

Proof of Theorem 2.2.1:

(a) Note that $r_{i1} = b_{i1}b_{11}$ and $b_{11} = 1$. Thus, $r_{i1} = b_{i1}$ for $i = 2, 3, ..., k$.

For $i = j$, $r_{ii} = 1 = \sum_{u=1}^{i} b_{iu}b_{iu}$. Therefore, $b_{ii} = \sqrt{1 - \sum_{u=1}^{i-1} b^2_{iu}}$.

To prove the form of b_{ij} for $2 \leq j < i \leq k$ in (2), we resort to induction on dimension (k) of R, where induction step starts from $k = 3$.

In this case, $r_{32} = b_{31}b_{21} + b_{32}b_{22}$. Hence, $b_{32} = \frac{r_{32} - b_{31}b_{21}}{b_{22}} = \frac{r_{32} - r_{31}r_{21}}{\sqrt{1-r^2_{21}}} = \rho_{32:1}\sqrt{1 - r^2_{31}}$, which satisfies (2).

Suppose (2) holds for any correlation matrix of dimension k. We require to prove it for a $(k+1)$ dimensional correlation matrix R_{k+1}. We write

$$R_{k+1} = \begin{bmatrix} R_k & r_{k+1} \\ r_{k+1} & 1 \end{bmatrix}$$

where R_k is the k-dimensional correlation matrix pertaining to first k variables and $r_{k+1}^\top = (r_{k+1,1}, r_{k+1,2}, ..., r_{k+1,k})$

Let $R_{k+1} = B_{k+1}B_{k+1}^\top$ be the Cholesky decomposition of R_{k+1}, where B_{k+1} is a lower triangular matrix. Then $B_{k+1}^\top = [B_k^\top : \lambda_{k+1}^\top]^\top$, where B_k is the lower triangular Cholesky factor of R_k. In the context of induction hypothesis, we only require to show that elements of λ_{k+1} are in the form of (2), where we assume elements of B_k are already in the form of (2). We note that $\lambda_{k+1,1} = r_{k+1,1}$ and equating $r_{k+1,j}$ from the Cholesky decomposition one gets,

$$\lambda_{k+1,j} = \frac{r_{k+1,j} - r_{j1}r_{i1} - \sum_{u=1}^{j-1} b_{ju}\lambda_{k+1,u}}{\sqrt{1 - \sum_{u=1}^{j-1} b_{ju}^2}} \tag{A.4}$$

$$= \frac{r_{k+1,j} - r_{j1}r_{i1} - \sum_{u=1}^{j-1} \rho_{uj:1,2,...,u-1}\rho_{u,k+1:1,2,...,m-1} \prod_{m=1}^{u-1} \sqrt{(1 - \rho^2_{m,k+1:1,2,...,m-1})(1 - \rho^2_{mj:1,2,...,m-1})}}{\prod_{u=1}^{j-1} \sqrt{1 - \rho^2_{uj:1,2,...,u-1}}} \tag{A.5}$$

$$= \frac{\rho_{j,k+1:1,2,...,j-1} \prod_{u=1}^{j-1} \sqrt{(1 - \rho^2_{uj:1,2,...,u-1})(1 - \rho^2_{u,k+1:1,2,...,u-1})}}{\prod_{u=1}^{j-1} \sqrt{1 - \rho^2_{uj:1,2,...,u-1}}} \tag{A.6}$$

$$= \rho_{j,k+1:1,2,...,j-1} \prod_{u=1}^{j-1} \sqrt{1 - \rho^2_{u,k+1:1,2,...,u-1}} \tag{A.7}$$

where (31) follows from induction hypothesis and Proposition 7.2 and (32) follows from Proposition 7.1.

(b) Since the diagonal entries of the matrix B are non-negative, the statement follows from (2.5) and the uniqueness of the Cholesky factor of R. Consequently, the angles θ_{ij}'s are simply the inverse cosine of the semi-partial correlations, and as such they are readily interpretable statistically.

A.2 Characterization of Compound symmetric structure

We consider the Cholesky decomposition of $R = BB^\top$, where we denote the $(i,j)^{th}$ entry of B by b_{ij}. For $i > j$, one can write $r_{ij} = \sum_{l=1}^{j} b_{il}b_{jl}$. For $j = 1$, $r_{i1} = b_{11}b_{i1}$, hence $b_{i1} = r$, since $b_{11} = 1$. From the relationship between angles and Cholesky factor, it follows that $\cos(\theta_{i1}) = b_{i1} = r$ for $j = 2, 3, ..., k$. Thus θ_{i1}'s are all equal to a common θ.

For any $i > j$, the proof follows by induction. For $i > j$, $r_{ij} = \sum_{l=1}^{j-1} b_{il}b_{jl} + b_{jj}b_{ij}$. By induction hypothesis, all the preceding angles and Cholesky factors are functions of r. Thus, the first term is a function of r. For the second term, we note $b_{jj} = \prod_{l=1}^{j-1} \sin(\theta_{jl})$, which involves all the preceding angles and thus a function of r and $b_{ij} = \cos(\theta_{ij}) \prod_{l=1}^{j-1} \sin(\theta_{jl})$. Thus it follows that $\cos(\theta_{ij})$ is a function of r and θ_{ij} is a function of θ. Therefore, the entire Θ matrix is characterized by a single angle θ.

In particular, the first column of Θ equals to θ. The second column is $\theta_{i2} = \arccos\left(\frac{\sqrt{r(1-r)}}{\sqrt{1+r}\sin\theta}\right)$ with $r = \cos\theta$ for $i = 3, 4, ..., k$. Similarly for third row and onward, one can solve for θ_{ij} which are explicit functions of θ. We denote θ as the pivotal angle which is all the entries in the 1^{st} column of Θ and all other entries are implied angles.

A.3 Characterization of AR(1) structure

We note that $\theta_{21} = \theta$ and other entries of first column of Θ is related to θ by the relation $\cos\theta_{i1} = r_{i1} = (\cos\theta)^{i-1}$ for $j = 3, 4, ..., k$. One can solve for other entries of Θ starting from second row onwards to derive the exact functional relationship with θ. Thus, we have one pivotal angle θ which is located at the $(2,1)^{th}$ entry of Θ and the remaining angles are implied angles.

A.4 Proof of Proposition 3.1.1

Proof. The proof follows using the Cholesky decomposition and relating it to Θ. Consider $R = BB^\top$ where $B = ((b_{ij}))$ as given in (2.5). Suppose Θ satisfies $\theta_{ij} = \pi/2$ for $|i-j| > \lambda$. Then from (2.5), it follows that $b_{ij} = 0$ for $|i-j| > \lambda$.

Thus, for $|i-j| > \lambda$, we have

$$r_{ij} = \sum_{l=1}^{k} b_{il} b_{jl} = \sum_{l=1}^{j} b_{il} b_{jl} \quad \text{(Assuming } i > j\text{)}$$

Since $b_{ij} = 0$ for $|i-j| > \lambda$ and l runs over $1, 2, ..., j$ and $i > j$, i.e. contribution of b_{jl} for each summand is 0 which implies $r_{ij} = 0$ for $|i-j| > \lambda$.

For the only if part, one has $r_{ij} = 0$ for $|i-j| > \lambda$ where $1 \leq \lambda < k$ and the proof follows by induction. To see this, consider,

$$r_{i1} = b_{i1} b_{11}, \qquad \text{(since } B \text{ is lower triangular matrix)}$$

By the construction of B, the diagonal entries are positive. Thus $r_{i1} = 0$ implies $b_{i1} = 0$ which in turn implies $\theta_{i1} = \pi/2$. For the induction step, assuming $j < i$ and $i - j > \lambda$, consider

$$\begin{aligned} r_{ij} &= \sum_{l=1}^{j} b_{il} b_{jl} \\ &= b_{ij} b_{jj} + \sum_{l=1}^{j-1} b_{il} b_{jl} \end{aligned}$$

From the induction hypothesis, $\theta_{jl} = \pi/2$ for $l \leq j-1$ implies the second summand is 0. Thus, we must have $b_{ij} = 0$ which implies $\theta_{ij} = \pi/2$ and this completes the proof. □

A.5 Proof of Proposition 3.3.1

Proof. The proof uses the Cholesky decomposition of R and the fact that Cholesky factor of a block diagonal matrix is also block diagonal and vice versa.

Hence, the lower triangular Cholesky factor R has the form $B = \text{block diag}(B_1, B_2, ..., B_m)$,

where B_i is an upper triangular Cholesky factor of R_i.

The proof will be complete if we can show that Cholesky factor B of a compound symmetric correlation matrix R can be written in terms of only one angle. From the relationship between angles and Cholesky factor, it follows that $\cos(\theta_{i1}) = b_{i1} = r$ for $j = 2, 3, ..., k$. Thus θ_{i1}'s are all equal to a common θ. For any $i > j$, the proof follows by induction. For $i > j$, $r_{ij} = \sum_{l=1}^{j-1} b_{il}b_{jl} + b_{jj}b_{ij}$. By induction hypothesis, all the preceding angles and Cholesky factors are functions of r. Thus, the first term is a function of r. For the second term, we note $b_{jj} = \prod_{l=1}^{j-1} \sin(\theta_{jl})$, which involves all the preceding angles and thus a function of r and $b_{ij} = \cos(\theta_{ij}) \prod_{l=1}^{j-1} \sin(\theta_{jl})$. Thus it follows that $\cos(\theta_{ij})$ is a function of r and θ_{ij} is a function of θ.

□

A.6 Proof of Proposition 3.3.2

Proof. First consider $|\theta_1 - \theta_2| = \delta$. Note that $|r_1 - r_2| = |\int_{\theta_1}^{\theta_2} \sin x dx|$. Also it is clear that $|r_1 - r_2|$ is an increasing function of $|\theta_1 - \theta_2|$, since sin is positive in $[0, \pi)$. Now since sin is increasing in $[0, \pi/2]$ and decreasing in $(\pi/2, \pi)$, $|r_1 - r_2|$ will take minimum value for $|\theta_1 - \theta_2| = \delta$ when $\theta_1 = 0, \theta_2 = \delta$. Thus the minimum value of $|r_1 - r_2|$ is $|1 - \cos\delta|$. □